

DECONSTRUCTING PHYLOGENETIC RECONSTRUCTION: EFFECTS OF  
ASSUMPTION VIOLATIONS ON EVOLUTIONARY INFERENCE

---

by

BRYAN KOLACZKOWSKI

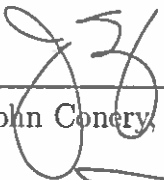
A DISSERTATION



Presented to the Department of Computer  
and Information Science  
and the Graduate School of the University of Oregon  
in partial fulfillment of the requirements  
for the degree of  
Doctor of Philosophy

December 2006

"Deconstructing Phylogenetic Reconstruction: Effects of Assumption Violations on Evolutionary Inference," a dissertation prepared by Bryan Kolaczkowski in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Computer and Information Science. This dissertation has been approved and accepted by:

  
\_\_\_\_\_  
Dr. John Conery, Co-chair of the Examining Committee

  
\_\_\_\_\_  
Dr. Joseph W. Thornton, Co-chair of the Examining Committee

27-Nov-2006  
Date

Committee in charge:

Dr. John Conery, Co-chair  
Dr. Joseph W. Thornton, Co-chair  
Dr. Michal Young  
Dr. Patrick Phillips

Accepted by:

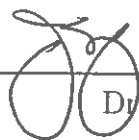
  
\_\_\_\_\_  
Dean of the Graduate School

Copyright 2006 Bryan Kolaczowski

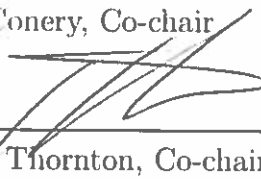
An Abstract of the Dissertation of  
Bryan Kolaczowski for the degree of Doctor of Philosophy  
in the Department of Computer and Information Science  
to be taken December 2006

Title: DECONSTRUCTING PHYLOGENETIC RECONSTRUCTION:  
EFFECTS OF ASSUMPTION VIOLATIONS ON EVOLUTIONARY  
INFERENCE

Approved:



Dr. John Conery, Co-chair



Dr. Joseph W. Thornton, Co-chair

Knowing how organisms are related evolutionarily is crucial for interpreting nearly all biological results. Evolutionary history is inferred using computational techniques that make simplifying assumptions about the evolutionary process. There is ample biological evidence that many of these assumptions are routinely violated, but little is known about the effects of assumption violations on phylogenetic inference.

Here I show how site-specific changes in evolutionary rates—an important evolutionary feature not incorporated into phylogenetic models—can cause existing methods to produce incorrect results. I develop a mixed branch length technique that produces more reliable inferences under realistic conditions. I outline a strategy to reduce the computational demands of the mixed branch length model by code opti-

mization and algorithm improvements.

Biologists also want to assess the confidence they should have in inferred phylogenies. Bayesian methods calculate posterior probabilities—i.e. the probability that a hypothesis is correct given the data, model, and prior probability distributions over model parameters—for phylogenetic hypotheses, producing an intuitively meaningful measure of statistical confidence, but concerns that posterior probabilities may regularly be too high has hampered acceptance of phylogenies produced using Bayesian methods. Understanding if, when, and why posterior probabilities are inflated is a crucial problem.

Here I show that although posterior probabilities are by definition correct assessments of subjective confidence given prior assumptions, they are accurate statements of objective confidence only when branch lengths are known in advance. When branch lengths are unknown, posterior probabilities can be either higher or lower than the long-run chance a hypothesis is correct. Posterior probabilities reported on actual phylogenies should therefore be interpreted only from a subjectivist standpoint.

My results suggest that phylogenetic techniques can produce incorrect phylogenies and assessments of statistical confidence due to assumption violations. Incorporating knowledge of how evolution works at the biological level into phylogenetic models can improve the quality of evolutionary inferences. The mixed branch length model incorporates an important feature of molecular evolution, potentially generating more accurate phylogenies than existing techniques.

This dissertation includes both my previously published and my co-authored materials.

## CURRICULUM VITAE

NAME OF AUTHOR: Bryan Kolaczowski

PLACE OF BIRTH: LaJolla, CA, U.S.A.

DATE OF BIRTH: November 6, 1974

## GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon  
Colorado State University  
University of New Mexico

## DEGREES AWARDED:

Doctor of Philosophy in Computer and Information Science, 2006,  
University of Oregon

Bachelor of Science in Computer and Information Science, 2001,  
University of Oregon

Bachelor of Science in Horticulture, 1996, Colorado State University

## AREAS OF SPECIAL INTEREST:

Bioinformatics  
Phylogenetics  
Statistical Inference  
Molecular Evolution

## PUBLICATIONS:

B. KOLACZKOWSKI AND J. W. THORNTON, *Is there a star tree paradox?* *Molecular Biology and Evolution*, 23 (2006), pp. 1819-1823.

J. W. THORNTON AND B. KOLACZKOWSKI, *No magic pill for phylogenetic error*, *Trends in Genetics*, 21 (2005), pp. 310-311.

B. KOLACZKOWSKI AND J. W. THORNTON, *Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous*, *Nature*, 431 (2004), pp. 980-984.

## ACKNOWLEDGMENTS

All the members of my committee have helped to shape and guide this work; for this I am grateful. John Conery opened the door to my study of computational science, allowed me the space to pursue an unconventional course of study, and constantly provided fresh points of view. Joe Thornton showed me how phylogenetics could be both interesting and important and helped me bridge the gap between computer science and biology.

My family has given me the impetus to strive to produce quality scientific work, consistently supported both my professional and personal goals, and dragged me away from work enough to keep me relatively sane.

I am additionally grateful to the faculty and students involved with the University of Oregon IGERT program in evo-devo. I am especially indebted to Drs. John Postlethwait and Patrick Phillips, without whom I would have completed my doctoral degree on a much tighter budget.

Thanks to Star Holmberg and the entire CIS department administrative staff, who helped me navigate the treacherous administrative hurdles standing between a BS and a PhD.

Patrick Phillips and Steve Proulx provided helpful comments and discussion regarding the analysis of heterotachy's effects on phylogenetic inference and the interpretation of Bayesian posterior probabilities. Helpful comments regarding the heterotachy experiments were also provided Robert DeSalle. Ziheng Yang and Paul Lewis helped me to better understand the star tree paradox, and Iain Pardoe kindly lent me his expertise on Bayesian statistics.

Support for this work was provided by an NSF IGERT training grant in Evolution, Development, and Genomics to the University of Oregon (NSF IGERT DGE-9972830) as well as an NSF grant to John Conery and Joe Thornton (NSF DEB-0516530).



To everyone working to bridge the gaps that separate us.

---

## TABLE OF CONTENTS

Chapter	Page
I. INTRODUCTION . . . . .	1
1.1 Model-Based Phylogenetic Inference . . . . .	3
1.1.1 Relative Transition Rate Models . . . . .	4
1.1.2 Evolutionary Rate Models . . . . .	7
1.1.3 Calculating Transition Probabilities . . . . .	8
1.2 Model Violations . . . . .	9
1.3 What is “Heterotachy”? . . . . .	10
1.4 Does Heterotachy Matter for Phylogenetic Inference? . . . . .	13
1.5 Bayesian Phylogenetic Inference . . . . .	16
II. PERFORMANCE OF MAXIMUM PARSIMONY AND LIKELIHOOD PHYLOGENETICS WHEN EVOLUTION IS HETEROGENEOUS . . . . .	20
2.1 Introduction . . . . .	20
2.2 Methods . . . . .	21
2.2.1 Simulations . . . . .	22
2.2.2 Phylogenetic Analyses . . . . .	23
2.2.3 Accuracy . . . . .	24
2.2.4 Bias and Error . . . . .	24
2.3 Results . . . . .	26
2.4 Discussion . . . . .	36
III. EFFECTS OF HETEROTACHY ON STANDARD AND MIXED BRANCH LENGTH PHYLOGENETIC INFERENCE . . . . .	39
3.1 Introduction . . . . .	39
3.2 Methods . . . . .	42
3.2.1 Phylogenetic Analysis . . . . .	42
3.2.2 Branch Length Heterogeneity . . . . .	43
3.2.3 Elongation Factor $1\alpha$ Sequences . . . . .	46
3.3 Results . . . . .	47

Chapter	Page
3.3.1 Branch Length Heterogeneity . . . . .	48
3.3.2 Elongation Factor $1\alpha$ Sequences . . . . .	59
3.4 Discussion . . . . .	63
<hr/>	
IV. OPTIMIZATION OF MIXED BRANCH LENGTH MODELS . . . . .	66
4.1 Computational Challenges in Phylogenetic Inference . . . . .	66
4.2 Why Simulated Annealing? . . . . .	69
4.3 Code Optimization . . . . .	70
4.3.1 Elongation Factor $1\alpha$ ( $ef1\alpha$ ) Data Profiling . . . . .	71
4.3.2 Dinoflagellate Data Profiling . . . . .	72
4.3.3 Bilaterian Data Profiling . . . . .	73
4.3.4 Planned Optimizations . . . . .	73
4.4 Faster Simulated Annealing . . . . .	74
4.5 Model Selection Using Simulated Annealing and Akaike Information Criterion (AIC) . . . . .	75
4.6 Parallel Algorithms . . . . .	77
4.7 Conclusion . . . . .	78
V. IS THERE A STAR TREE PARADOX? . . . . .	80
5.1 Introduction . . . . .	80
5.2 Methods . . . . .	82
5.3 Results . . . . .	83
5.4 Discussion . . . . .	89
VI. EFFECTS OF PRIOR BRANCH LENGTH UNCERTAINTY ON BAYESIAN POSTERIOR PROBABILITIES FOR PHYLOGENETIC HYPOTHESES . . . . .	90
6.1 Introduction . . . . .	90
6.2 Methods . . . . .	95
6.2.1 Bayesian Analyses . . . . .	95
6.2.2 Accuracy of BMCMC . . . . .	95
6.2.3 Simulations . . . . .	96
6.2.4 Comparing Posterior Probabilities . . . . .	98

Chapter	Page
6.3 Results . . . . .	98
6.3.1 Accuracy of BMCMC . . . . .	98
6.3.2 Uncertainty Affects Posterior Probabilities . . . . .	99
6.3.3 The Effects of Branch Length Uncertainty on Posterior Probabilities Are Determined by the Pattern of Branch Lengths on the True Tree . . . . .	105
6.3.4 Even Very Long Sequences Do Not Eliminate the Effects of Branch Length Uncertainty . . . . .	112
6.3.5 Diffuse Priors Are Preferable to Small-Mean Exponential Priors . . . . .	114
6.4 Discussion . . . . .	116
VII. CONCLUSION . . . . .	120
BIBLIOGRAPHY . . . . .	124

## LIST OF FIGURES

Figure	Page
1.1 Main components of phylogenetic models of evolution. . . . .	6
1.2 Types of evolutionary rate variation in molecular sequence data. . . . .	12
2.1 Likelihood-based methods are less accurate than maximum parsimony under heterogeneous conditions. . . . .	27
2.2 Maximum parsimony outperforms likelihood-based methods when strong support is required. . . . .	28
2.3 The BL <sub>50</sub> defines a method's inconsistency point. . . . .	29
2.4 Evolutionary heterogeneity biases likelihood-based methods. . . . .	30
2.5 Parsimony outperforms likelihood over a wide range of heterotachous conditions. . . . .	32
2.6 Maximum parsimony is more accurate than likelihood methods when techniques to improve phylogenetic performance are used. . . . .	33
2.7 Maximum likelihood fails to correctly infer the lengths of internal and terminal branches from heterotachous data. . . . .	34
2.8 Poor maximum likelihood performance is due to assuming homogeneous branch lengths. . . . .	35
2.9 Violating the identical distribution assumption causes likelihood-based methods to be statistically inconsistent when the correct heterotachous evolutionary model is used. . . . .	36
3.1 Summary of branch length heterogeneity simulations. . . . .	45
3.2 Accuracy of standard model-based phylogenetic inference is impaired by Felsenstein Zone Heterotachy; the mixed branch length model improves accuracy. . . . .	50
3.3 Standard phylogenetic techniques are strongly biased in favor of the correct tree under Inverse Felsenstein Zone Heterotachy; the mixed branch length model is unbiased. . . . .	53
3.4 Single Long Branch Heterotachy impairs the accuracy of standard model-based phylogenetic techniques; the mixed branch length model improves accuracy. . . . .	55
3.5 Standard maximum likelihood misestimates branch lengths under Single Long Branch Heterotachy; the mixed branch length model provides improved estimates. . . . .	57
3.6 Signal-Noise Heterotachy (SNH) impairs the accuracy of phylogenetic inference. . . . .	58

Figure	Page
3.7 A mixed branch length model recovers the correct Microsporidia + Fungi grouping from elongation factor $1\alpha$ sequence data. . . . .	60
3.8 Mixed model analysis of elongation factor $1\alpha$ data partitions sites into branch length categories. . . . .	62
4.1 Runtime profiling reveals data structure copying as the major performance bottleneck. . . . .	72
5.1 Variance in posterior probability of a resolved tree does not increase with increasing sequence length. . . . .	84
5.2 Type I error rates based on posterior probability are conservative. . . . .	85
5.3 Star-tree generated data with ideal character state pattern frequencies produce equal posterior probability for each possible resolved tree. . . . .	88
6.1 BMCMC accurately estimates Bayesian posterior probabilities. . . . .	100
6.2 Branch length uncertainty affects posterior probabilities. . . . .	103
6.3 Branch length patterns affect posterior probabilities when few terminal branches are long. . . . .	106
6.4 Branch length patterns affect posterior probabilities when many terminal branches are long. . . . .	107
6.5 Branch length patterns affect posterior probabilities when two terminal branches are long and the other two are short. . . . .	110
6.6 Branch length uncertainty affects posterior probabilities under empirically derived conditions. . . . .	113
6.7 Small-mean exponential branch length priors produce posterior probabilities that deviate more strongly from those produced using the true prior distribution than uniform priors. . . . .	115

## CHAPTER I

### INTRODUCTION

Empirical data by themselves are typically not very interesting; rather, the inferences we draw from the data are what give data meaning. Inferential techniques produce general information about the hidden processes that govern the world in which we live from specific data instances, but all techniques that infer information from empirical data make assumptions about the underlying processes that generated the data. These assumptions can be made because 1) we believe the data-generating process to have certain properties, or 2) we require certain simplifications in order to perform calculations or otherwise extract information from the data. In either case, what we assume in order to make inferences often turns out to be wrong, potentially undermining the quality of the resulting information. Understanding which of our assumptions are incorrect, as well as the effects of assumption violations on the accuracy of our inferences, is therefore crucial for reliably interpreting inferred information. In addition, a thorough understanding of how assumption violations affect resulting inferences can lead to the development of novel techniques that produce more accurate results under realistic conditions.

Phylogenetic inference is a computational technique for reconstructing the evolutionary history of living organisms or genes (usually depicted as a bifurcating tree) from molecular sequence data (typically either DNA or protein sequences) or other characters. The centrality of phylogenetic inference to evolutionary biology is unquestionable; the only figure in Darwin's seminal work, *The Origin of Species*, is a tree-like structure depicting the historical relationships among present day

organisms [13]. Because nearly all biological results are meaningful only in the context of evolution [15], phylogenetic inference is fast becoming one of the fundamental computational techniques used throughout modern biology as a whole. Phylogenetic inference provides a necessary framework for all valid comparative biology [40] and has dramatically enhanced our knowledge of the epidemiology of infectious disease [25, 34] and the complex interactions underlying ecological processes [122, 125].

Phylogenetic inference is a unique and challenging problem for a number of reasons. First, because evolution proceeds over immense time scales and leaves little evidence of its process, the true phylogeny is long past and can never be known with certainty. Fossil evidence is scarce, difficult to interpret, and cannot provide clear molecular data about the distant past. Second, because characters that appear very similar can have independent evolutionary origins (a phenomenon called “convergent evolution”), misleading phylogenetic information is not only possible but can sometimes overwhelm the true evolutionary signal in cases of ancient evolutionary radiations and other difficult problems. Finally, the ways in which species diversify and evolve over time are governed in large part by external processes that are difficult to predict or model mathematically, such as changes in global or local climate, cataclysmic geological events, interactions with other organisms, and population-level dynamics. As a result, models used to infer phylogenies—which typically include only molecular-scale dynamics—are necessarily simplified compared to the true evolutionary process.

While little can be done to augment the paucity of data from extinct and ancestral organisms, likelihood models have been developed to formally evaluate molecular convergence, and models that more closely approximate complex real-world evolutionary dynamics are constantly being developed. Nonetheless, all phylogenetic methods make simplifying assumptions about the process of molecular evolution, many of which can be regularly violated by empirical sequence data. Most techniques accurately reconstruct evolutionary relationships when their assumptions are met, but violating the assumptions of a method can result in inaccurate inferences. For example, maximum parsimony (MP, a simple



nonparametric technique) assumes that convergent evolution is not structured to favor a particular tree. When evolutionary rates are significantly accelerated in non-sister lineages, this assumption can be strongly violated, producing a topological bias in favor of placing long branches together on the tree regardless of the actual evolutionary history [21]. Although significantly more sophisticated than simple MP, model-based phylogenetic methods also rely on a simplified model of the evolutionary process and therefore may be prone to similar errors.

## 1.1 Model-Based Phylogenetic Inference

Model-based phylogenetics such as maximum likelihood (ML) and Bayesian Markov Chain Monte Carlo (BMCMC) employ an explicit probabilistic model of the evolutionary process that can account for accelerated evolutionary rates in different lineages, producing highly accurate reconstructions under conditions that confound maximum parsimony [115, 116, 127]. In addition, because different evolutionary models can be specified and evaluated using well-founded statistical methodology [88, 90], it is possible to formulate and test hypotheses about the process of molecular evolution—not just the tree topology—using model-based techniques [52]. There is hope that because the evolutionary model can be made arbitrarily complex, as important features of the molecular evolutionary process are identified, these can be incorporated into the model, resulting in increasingly accurate inferences.

Given an evolutionary model, it is possible to infer a phylogeny that is ‘optimal’ under that model using the likelihood principle. If the tree topology ( $t$ ) and all free parameters of the model  $M$  are known, the probability of any possible sequences  $X$  occurring at the tips of the tree ( $P(X|t, M)$ ) can be calculated directly from the model. Reversing this process—taking the sequence data as given—the ‘likelihood’ of the model can be calculated under any set of parameter values using the same formula:  $L(t, M|X) = P(X|t, M)$  [18]. Parameters can be optimized by choosing values that maximize the model’s likelihood (the ML approach), or likelihoods can be calculated by integrating over multiple parameter values using Bayesian Markov Chain Monte Carlo.

Although it is possible to calculate likelihoods using any probabilistic model of sequence change on a tree, all current phylogenetic models are based on a continuous-time Markov process. Given a finite number of possible states—A,C,G,T in the case of nucleotide data, the 20 amino acids in the case of protein data—a Markov model specifies the conditional probability that, if a character is in state  $i$  at time  $u$ , the character will be in state  $j$  at time  $u + v$ . In a phylogenetic context, ‘time’ is measured as the expected number of substitutions/site along each branch on the tree (i.e. the ‘branch length’), and the likelihood for each site can be efficiently calculated using a tree-traversal algorithm [22]. Under the standard assumptions that the state at one site is not influenced by the states at other sites (independence) and that all sites follow the same evolutionary model with the same parameter values (identical distribution), the total likelihood for an entire molecular sequence alignment is simply the product of all individual site-likelihoods.

One of the benefits of likelihood phylogenetics is that information about the molecular evolutionary process can be formalized and incorporated into the evolutionary model, potentially improving the accuracy of resulting phylogenetic inferences. Common models used today have two main components: a component describing the instantaneous transition rate from any state to any other state and a component describing the rate of evolution.

### 1.1.1 Relative Transition Rate Models

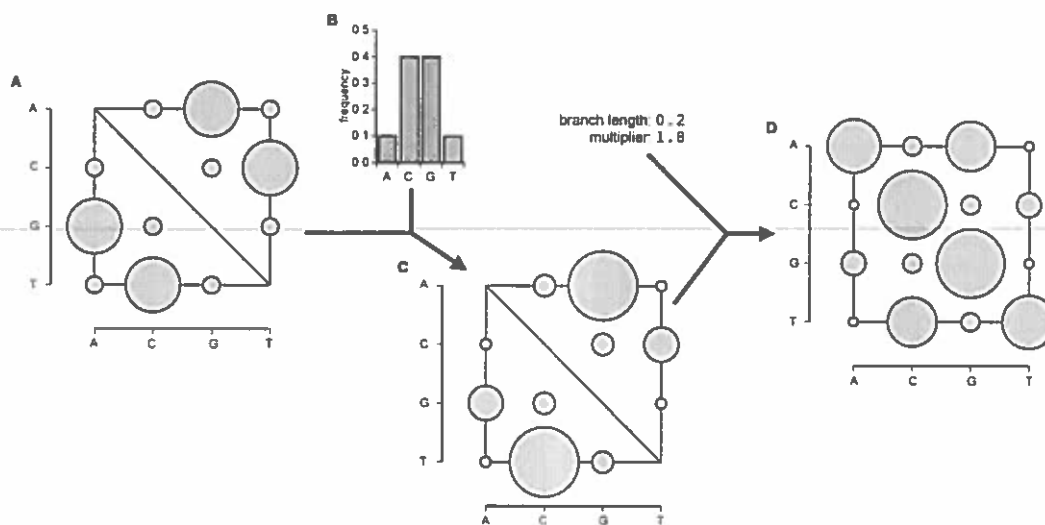
Relative transition rates are modeled by extracting two pieces of information from the data: a ‘substitutability’ matrix ( $S$ -matrix) describing the propensity for different states to substitute for one another and a frequency vector ( $\pi$ ) describing how often each state appears in the molecular sequence data under study. Each of these pieces making up the relative transition rate model have been developed to explain observed features of molecular evolution.

### Substitutability Matrix (*S*-matrix)

First, it has been observed that different types of state changes may be more or less frequent than others. In the case of nucleotide data, transitions ( $A \leftrightarrow G, C \leftrightarrow T$ ) are typically more common than transversions [59]. In the case of protein data, it is reasoned that because the molecular interactions upon which life depends are determined in large part by the biochemical properties of the molecules involved, changes preserving the biochemical properties of the amino acid at a given position should tend to occur more frequently than changes that radically alter a site's biochemistry. For example, changes among hydrophobic residues are more frequent than changing a hydrophobic amino acid to a hydrophilic one. The relative substitutability among the various possible states is expressed by an  $N \times N$  matrix (where  $N$  is the number of possible states) called an *S*-matrix. Each entry ( $i, j$ ) in the *S*-matrix indicates the relative 'substitutability' among states  $i$  and  $j$ , with large values indicating substitutions that are likely to occur and small values indicating rare substitutions (Figure 1.1A). To maintain computational tractability, it is assumed that the *S*-matrix is symmetrical or 'time-reversible': the relative transition rate from  $i \rightarrow j$  is always the same as the rate from  $j \rightarrow i$ . Matrices of various complexity can be used, ranging from very simple models that assume all changes have equal weight to a general time-reversible model in which the relative substitutability between each pair of states is a separate free parameter. In the case of protein data—where the large number of possible states makes parameter estimation more difficult—empirical *S*-matrices have been developed by analyzing large numbers of protein sequences and calculating the 'average' substitutability matrix over these sequence data [57, 123]. Empirically-derived matrices—which have no free parameters—can then be applied to novel phylogenetic problems.

### State Frequency Vector ( $\pi$ )

The state frequency vector ( $\pi$ ) is used to augment the information in the *S*-matrix to account for the observation that different states typically occur with different frequencies in molecular sequence data [22]. For example, different genomic



**FIGURE 1.1:** Main components of phylogenetic models of evolution. (A) Substitutibility matrix ( $S$ -matrix) describing the relative substitutibility of nucleotides. Large bubbles indicate nucleotides that substitute for one another more readily than small bubbles; this example shows a 10:1 transition bias. (B) State frequency vector ( $\pi$ ) describing the frequency of each nucleotide in the data set. This example shows elevated frequencies of G and C. (C) Each entry in the  $S$ -matrix is multiplied by the appropriate state frequency to produce a  $Q$ -matrix. (D) The  $Q$ -matrix is first multiplied by the ASRV-corrected branch length then exponentiated to produce the transition probability matrix ( $P$ -matrix) which gives the probability of each possible nucleotide change along the branch.

regions regularly vary in Guanine+Cytosine (GC) content [8]; some regions are GC rich, while others are relatively GC poor. Each entry in the matrix  $S_{i \rightarrow j}$  is multiplied by the frequency with which state  $j$  occurs in the sequence data in order to reflect the idea that changing to a more frequent state is more likely than changing to an infrequent state (Figure 1.1B,C). Typically, the frequency of each state is estimated directly from the data as a free parameter, and state frequencies are assumed to be ‘stationary;’ that is, state frequencies are assumed to remain constant over evolutionary time. The result of combining the  $S$ -matrix and the frequency vector is a new matrix called an  $R$ -matrix that describes the relative instantaneous transition rates among any two possible states.

### 1.1.2 Evolutionary Rate Models

The second main component of modern evolutionary models used for phylogenetic inference is a model of the rate of evolution. As with relative transition rates, evolutionary rate models typically employed today have two pieces: a piece describing lineage-specific evolutionary rates and a piece describing among-site rate variation. Because it has been known for some time that the rate of evolution can be faster or slower in different lineages, current evolutionary models have a free 'branch length' parameter associated with each branch on the tree that describes the rate of evolution on that branch. Branch lengths—which are typically estimated from the data—are expressed in terms of the expected number of substitutions/site along the molecular sequence. Expressed in this way, branch lengths actually conflate two evolutionary parameters; the branch length is the rate of evolution along the branch times the amount of time between two speciation events. Interestingly, although adding taxa to the tree does not necessarily increase the complexity of the relative transition rate model, the addition of taxa must increase the complexity of the evolutionary rate model, because new branch length parameters are added to the tree.

#### Modeling Among-Site Rate Variation (ASRV)

The second piece of the model describing evolutionary rates is a model of among-site rate variation (ASRV). Because different sites in a molecular sequence are under different levels of evolutionary constraint—some sites are highly constrained and tend not to accumulate mutations while others are more variable—different sites can evolve at different rates, even within the same lineage [128]. When a single branch length is applied to sites evolving at different evolutionary rates, the branch length underestimates the rate of fast sites and overestimates the rate of slow sites, resulting in incorrect rates for most sites and very poor phylogenetic accuracy [127]. As it is typically unknown which sites evolve at which rates, it is not possible in general to assign sites to different rate categories a priori; instead, the likelihood at each site is calculated as a weighted average over

all possible rates. In an ASRV model, a pre-determined number of rate multipliers  $r = (r_1, r_2, \dots, r_n)$  either ‘stretch’ or ‘compress’ the tree’s branch lengths (by multiplying the raw branch length by the rate multiplier to get a new ASRV-corrected branch length) to account for different evolutionary rates at different sites. A proportion of sites  $p = (p_1, p_2, \dots, p_n)$  are estimated to evolve at each evolutionary rate, and the likelihood for a site is calculated as the weighted sum over all rate multipliers:  $L(X|t, M, r, p) = \sum_{i=1}^n p_i L(X|t, M, r_i)$ , where  $M$  is the evolutionary model and  $t$  is the tree topology with branch lengths. Of course, for branch lengths to retain their intended meaning—the expected number of substitutions/site—it is required that  $\sum_{i=1}^n p_i r_i = 1.0$ .

A very general ASRV model like the one just described has many parameters (each rate multiplier and site proportion must be estimated separately), and the constraints on the parameter values are complex, making optimization difficult. Very simple ASRV models that assume sites are either fixed (invariant) or variable have been proposed, but it is generally accepted that such a simple classification of sites into ‘on’ and ‘off’ categories is too gross to account for the subtle evolutionary dynamics affecting ASRV [42, 128]. The most common ASRV model used today—the “discrete gamma model”—attempts to strike a balance between these two extremes. The gamma ASRV model assumes a number of rate multipliers—typically 4–8 rather than just two in the ‘invariant sites’ model—but these multipliers are distributed according to a gamma distribution [126]. Using this model, a single parameter value—the shape of the gamma distribution—determines any number of rate multipliers, and the proportion of sites evolving at each rate is assumed to be equal. The discrete gamma model has been shown to produce accurate phylogenies under realistic conditions where other models fail [46, 63, 116, 131].

### 1.1.3 Calculating Transition Probabilities

Multiplying a branch length by a rate multiplier gives an ASRV-corrected branch length that can be used in conjunction with the relative transition rate matrix to

calculate the conditional probability of each state change along the branch. First, the  $R$ -matrix is scaled so that the sum of the off-diagonals is 1.0, producing a  $Q$ -matrix. The transition probability matrix  $P$  is then calculated by taking:  $e^{Qbr}$ , where  $b$  is the initial branch length, and  $r$  is the rate multiplier (Figure 1.1D). Each entry in the  $P$ -matrix ( $P_{i,j}$ ) gives the probability of changing from state  $i$  to state  $j$  along the given branch on the tree; the total probability of observing a set of states at the tips of the tree can be computed using a post-order tree traversal [22].

## 1.2 Model Violations

When the correct evolutionary model is used, maximum likelihood converges on the correct tree as sequence length increases (i.e. it is 'consistent'), and Bayesian methods using the same probabilistic models may be expected to perform similarly. When the model is very complex, however, different optimal trees can have the same likelihood—the tree is 'nonidentifiable'—and ML is no longer capable of distinguishing the correct tree from incorrect alternatives, even when the true model is used and infinite data are available [107]. Moreover, computer simulations have shown that when data are generated using a complex model but analyzed using a simpler 'underparameterized' model, incorrect inferences of tree topology, model parameters, and statistical support for phylogenetic hypotheses are regularly made [11, 53, 62, 65, 101, 104, 111]. So, if the correct evolutionary model is available—and the tree topology is identifiable under the model—we can be assured that likelihood methods have excellent asymptotic properties; however, this guarantee is void if the model is either very complex or incorrect.

Evolutionary models used to infer phylogenies typically have a number of free parameters that must be estimated from the sequence data. Failure to accurately estimate model parameters can result in phylogenetic error, even if the model is correct, but a certain degree of homogeneity must be assumed so that enough data are available to ensure accurate parameter estimates. If every site in the sequence evolved under unique dynamics, there would be no data to provide statistically meaningful parameter estimates for each site. With the notable exception of

among-site rate variation [128]—which incorporates faster- and slower-evolving sites—evolutionary models typically assume that all sites evolve under the same evolutionary dynamics. This maximizes the amount of data available to estimate the parameters of the model, but there is ample empirical evidence that different sites regularly evolve under different evolutionary constraints, resulting in a violation of the homogeneity assumption implicit in existing evolutionary models. In some cases, inferences can be robust to certain violations of the model’s assumptions [41, 110, 131]; other times assumption violations can cause incorrect trees to be strongly supported [9, 55, 62]. If model violations commonly found in empirical sequence data undermine the credibility of existing phylogenetic inference techniques, the potential repercussions for biological science are profound, as phylogenetic techniques are ubiquitously employed, and many results are dependent on evolutionary relationships being correct.

### 1.3 What is “Heterotachy”?

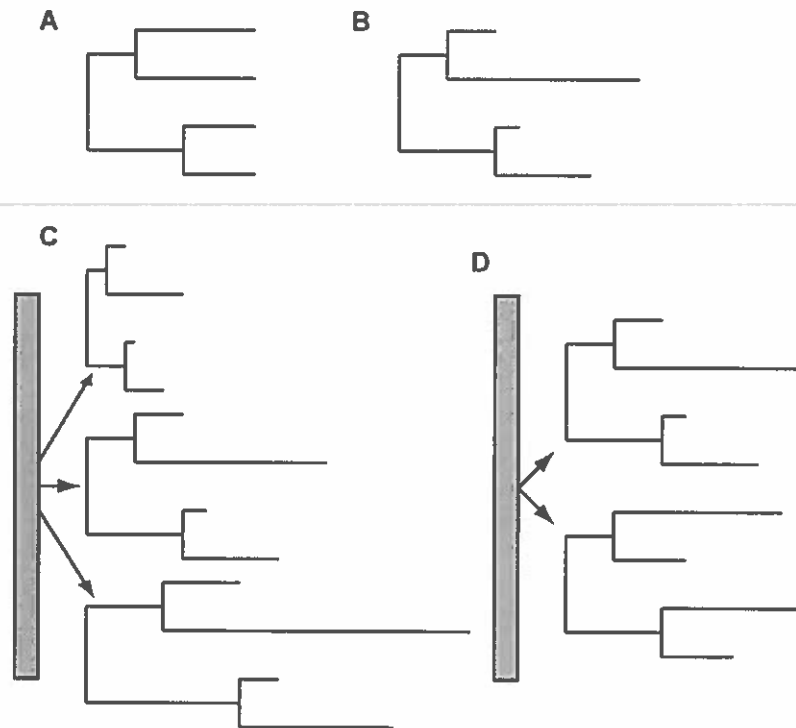
Both among-lineage and among-site rate variation are widely recognized as important evolutionary features. Among-lineage rate variation is incorporated into phylogenetic models by lineage-specific branch lengths, and among-site rate variation is typically modeled using a discrete gamma approach. A related molecular evolutionary feature that is not modeled by current phylogenetic techniques is “heterotachy,” or site-specific evolutionary rate variation [27, 48, 55, 69, 71, 72, 75, 76, 84, 87, 91]. Heterotachy was first identified as an important molecular evolutionary feature in the mid-1970’s by Walter Fitch and colleagues, who were studying the evolution of vertebrate cytochrome protein sequences [26, 27, 76]. These researchers noticed that the identities of the variable sites in the sequence were different in different lineages; in some lineages, a specific site was constant in all extant taxa, while in another lineage that same site was highly variable. Other sites exhibited the reversed pattern, being variable in the first lineage and constant in the second. The observation of this pattern led to the development of the “covarion” hypothesis of evolution, which states that at any



given time, only some of the sites in a sequence are capable of accepting substitutions—the remaining sites being fixed by selective constraints—but the specific sites that are capable of varying can change over time [27, 120]. Weak evidence for the covarion hypothesis has been generated by phylogenetic studies showing that a covarion model typically provides an improved statistical fit to empirical sequence data compared to simpler ‘homotachous’ models [48, 71]. A second early observation regarding the covarion model was that not only which sites are variable changes in different lineages, but how many sites are variable can also be different across phylogenetic lineages [27]. This early observation has recently been verified for additional molecular sequence data using more sophisticated statistical estimation procedures [69].

The recent explosion of molecular sequence data has allowed the development of more fine-grained statistical tests for site-specific evolutionary rate shifts, giving us a more detailed picture of what heterotachy may look like. Recent analyses have shown that—in addition to the ‘on’/‘off’ dynamics predicted by the covarion model—sites in a molecular sequence can regularly undergo more subtle evolutionary rate shifts in which a site can switch from evolving slowly to evolving quickly and vice versa [35, 36, 72, 84]. A current general model of heterotachy views this phenomenon as an interaction between lineage-specific and site-specific evolutionary rate variation (Figure 1.2). In among-lineage rate variation—modeled by lineage-specific branch lengths—different lineages can be fast- or slow-evolving, but all sites evolve at the same rate in all lineages. Under an ASRV model, different sites can be either fast or slow, but fast sites are fast in all lineages, and slow sites are always slow. Under heterotachous evolution, some sites evolve fast in some lineages and slower in other lineages, while other sites exhibit reversed evolutionary rates, evolving slowly in lineages where other sites evolve quickly and quickly where other sites evolve slowly. The result of heterotachy is a series of evolutionary rate shifts that can occur throughout the phylogenetic tree but may apply only to some positions in the sequence.

One of the functional explanations for heterotachy is that, as sequences diverge, new molecular functions can be acquired—and ancestral functions lost—causing



**FIGURE 1.2:** Types of evolutionary rate variation in molecular sequence data. (A) No rate variation: all sites evolve at the same evolutionary rate in all lineages. The total length from the root of the tree (the common ancestor of all taxa under study) to each tip is the same. (B) Among-lineage rate variation: sequences evolve at different rates in each lineage, but all sites evolve at the same relative rate. (C) Among-site rate variation (ASRV): different sites evolve at different evolutionary rates, but rates are proportional across sites. (D) Heterotachy, or site-specific rate variation, occurs when different sites evolve at different non-proportional rates.

unconstrained sites to become fixed in certain lineages and vice versa. As sites that are free to vary become involved in a new function, novel evolutionary constraints cause these sites to become slow-evolving [68, 69, 108]; conversely, loss of function in specific lineages can free previously constrained sites, allowing them to become more variable [55]. Sites involved in different molecular subfunctions could exhibit different evolutionary rate patterns on the tree due to various subfunctions arising in different lineages, leading to heterotachous evolution. Although functional divergence is one potential cause of heterotachy, heterotachous sites are not

typically linked to known functional or structural domains but seem to occur more-or-less evenly dispersed across the entire molecular sequence [72, 81]. This suggests that heterotachy may be a very general feature of molecular evolution and not necessarily linked to dramatic functional shifts.

A possible explanation for heterotachy without functional change is that nearby positions in a sequence may be able to perform similar molecular roles; heterotachy could arise as different positions become fixed to perform various roles in different lineages. For example, if a stabilizing intramolecular interaction improves the function of a protein, but either of two consecutive sites could perform the stabilizing role, different positions could be fixed to stabilize the molecule in different lineages. One of the sites may be stabilizing—and thus constrained—in some lineages but not in others, resulting in differential rates across lineages for that site. The neighboring site may then exhibit a complementary rate pattern, being variable in lineages where the first site is constrained and constrained where the first site is variable.

Heterotachy could therefore be caused either by shifts in molecular function or through random processes in which different sites in the molecular sequence are recruited to perform similar functions in different lineages. If these and other possible causes of heterotachy are important for molecular evolution, heterotachy could be very widespread and may in fact be the ‘rule’ rather than an interesting exception.

## 1.4 Does Heterotachy Matter for Phylogenetic Inference?

Even though the importance of heterotachy in molecular evolution has been well established, very little is known about the potential effects of heterotachy on phylogenetic inference. It has been predicted theoretically that when different sites evolve under different branch lengths—which is one way of modeling heterotachy—homotachous models may be statistically inconsistent [11], although

no proof has been given for this conjecture. An intriguing empirical study has suggested that heterotachy may be the chief cause of phylogenetic error when elongation factor 1 $\alpha$  (EF1 $\alpha$ ) data are used to reconstruct the Eukaryotic phylogeny [55]. When a homotachous model was used to reconstruct the Eukaryotic tree using EF1 $\alpha$  sequences, the spurious placement of the Microsporidia with the Archaeobacterial outgroup—rather than the correct placement of Microsporidia with Fungi—was weakly supported. When sites that exhibited a marked rate shift across the Archaeobacterial and Eukaryotic subtrees were removed from the analysis, support for the incorrect tree vs. the correct tree was significantly reduced, suggesting that these ‘heterotachous’ sites are at least in part responsible for the phylogenetic error. These results strongly suggest that heterotachy can be an important factor contributing to a reduction in phylogenetic accuracy, but many important questions remain unanswered.

First, the specific causes of heterotachy-induced phylogenetic error remain unclear. Presumably, the failure of homotachous models to correctly estimate site-specific evolutionary rates contributes to reduced accuracy under heterotachous conditions, but the mechanisms responsible for specific errors are not known. Understanding the precise causes of error can potentially generate information that can be used to develop new methods less prone to heterotachy-induced artifacts. Second, the potential types of errors likely to be seen when homotachous models are applied to heterotachous data is almost completely unknown. Heterotachy could cause certain trees to be preferred over others due to induced topological bias; branch lengths could be either over or underestimated, or heterotachy could cause reduced statistical power to resolve difficult phylogenies when homotachous models are applied. Understanding what erroneous inferences are likely to look like might help diagnose cases in which heterotachy may be causing an artifactual topology to be supported. Finally, the potential of new evolutionary models to improve the quality of phylogenetic inferences has not been adequately explored. The only available model that does include heterotachous dynamics is the ‘covarion’ model [29, 120], which incorporates a simplified form of heterotachy based on the covarion hypothesis and may not be adequate to capture more subtle evolutionary rate shifts.

Little is currently known about the relative accuracy of the covarion model vis-a-vis existing homotachous models. Other models of heterotachy are possible, but these have not been developed or tested prior to the work presented here.

The first part of this dissertation (chapters 2-4) addresses these questions. Chapters 2 and 3 examine the effects of heterotachy on model-based phylogenetic inference. Chapter 2 is an in-depth analysis of a single form of simulated heterotachy; I use a series of controlled simulation experiments to show how heterotachous evolutionary dynamics can severely impair the accuracy of phylogenies inferred using standard model-based methods (ML and BMCMC) by inducing a bias in favor of an incorrect phylogeny. I introduce a mixed branch length model that substantially improves phylogenetic accuracy under simulated conditions, potentially improving the quality of phylogenies inferred from actual sequence data. Chapter 3 examines additional forms of heterotachy using both simulated and empirical sequence data. I show that, in general, various forms of heterotachy can impair existing model-based techniques, resulting in incorrect inferences from both simulated and real-world data sets. I further develop and implement a general mixed branch length model for inferring phylogenies under heterotachous conditions. Applying this model to both simulated and empirical data, I show that incorporating heterotachy using a mixed branch length model can substantially improve the accuracy of phylogenetic inferences under realistic conditions. These results suggest that heterotachy is an important concern in phylogenetic inference, and that a mixed branch length approach is a useful tool for improving the quality of reconstructed phylogenies. Finally, chapter 4 addresses some of the computational issues raised by a complex mixed model approach. In particular, the mixed model requires much more computer time than simpler homotachous models, potentially limiting its applicability to large real-world data sets. Chapter 4 outlines a number of strategies for reducing the computational costs of mixed branch length analyses.

## 1.5 Bayesian Phylogenetic Inference

The popularity of Bayesian phylogenetic techniques using Markov Chain Monte Carlo (BMCMC) [54] has increased dramatically in recent years. Bayesian phylogenetics is appealing for at least three reasons: 1) the efficiency of the MCMC algorithm allows large phylogenies to be reconstructed from very large data sets using complex evolutionary models, all of which can improve the accuracy of phylogenetic reconstructions. 2) Bayesian techniques allow uncertainty in the values of model parameters to be incorporated into the analysis by integrating over a range of values, and 3) Bayesian posterior probabilities provide an intuitively meaningful measure of statistical confidence in phylogenetic hypotheses. In contrast, maximum likelihood (ML) analysis is computationally costly and provides no formal means of incorporating parameter uncertainty. Furthermore, the measure of statistical support used most often in ML analysis has been shown to be conservatively biased, resulting in reduced statistical power and a high rate of type II error [44].

Although Bayesian techniques have been used to resolve difficult and long-standing phylogenetic problems with strong statistical support [58, 79], there is growing concern that posterior probabilities on phylogenetic trees may regularly be too high, resulting in inflated confidence in uncertain results and a high rate of false inferences [12, 17, 67, 74, 102, 112, 117, 132]. Understanding whether posterior probabilities are inflated always or only under specific conditions—and why—is of crucial importance for interpreting the results of Bayesian analyses. This information can also be used to guide the development of new, more accurate statistical estimation techniques.

It has recently been suggested that not sampling unresolved trees as part of a Bayesian analysis could lead to overestimation of statistical confidence when the true tree is either multifurcating or has very short internal branch lengths [67, 132]. Existing Bayesian phylogenetics software samples only fully-resolved bifurcating phylogenies; multifurcating polytomies are approached by sampling very small branch lengths, but explicitly polytomous trees are not considered. When data are simulated on a small unresolved tree, Bayesian techniques produce equal posterior

probabilities for each possible resolved phylogeny when sequences are very short, but longer sequences occasionally produce strong support for one resolved tree over the others, even though there is no legitimate phylogenetic signal in the data. These results have led to predictions that posterior probabilities will become increasingly unpredictable as sequence length increases when data are simulated on an unresolved “star tree” [67, 132], although this prediction has not been explicitly tested. Previous studies have not examined whether high support for one resolved tree over the others occurs more frequently than expected due to stochastic error alone, and too few sequence lengths were examined to establish a general trend.

Another potential cause of erroneous confidence estimation in Bayesian phylogenetics is inaccurate prior information. In addition to the evolutionary model employed by both ML and BMCMC, Bayesian methods require the specification of prior probability distributions on all free model parameters. In effect, these prior distributions amount to additional assumptions required to conduct a Bayesian analysis; if these assumptions are wrong—as they are likely to be in any real analysis—BMCMC might produce biased phylogeny estimates or inaccurate assessments of statistical confidence. Unfortunately, very little is known about the effects of different prior assumptions on Bayesian phylogenetic techniques. It has recently been shown that when data are simulated with branch lengths drawn from an exponential distribution with specified mean, and the same distribution is used as a branch length prior in Bayesian analysis, the average posterior probability of a group of trees is the same as the proportion of inferred trees that are correct [53, 132]. However, when the prior mean on internal branch lengths is higher than the true mean, posterior probabilities are skewed upward; prior means lower than the true mean skew posterior probabilities downward [132].

These results establish that prior assumptions can affect posterior probabilities, but several important questions remain unanswered. First, real evolutionary history follows a single historically correct tree, whereas the simulations employed by previous authors [53, 132] generated phylogenetic trees and branch lengths using a stochastic process. How different prior assumptions affect posterior probabilities when there is a single correct tree with fixed branch lengths is unknown. Second,

previous studies examined the effects of various priors for the internal branch only, with the true prior distribution always assumed for terminal branches. Most empirical studies use the same prior distribution on both internal and terminal branches, and how different branch length priors applied across the entire tree affect posterior probabilities is unknown. Third, it has been common to use a uniform prior distribution with a large upper bound on branch lengths to represent prior ignorance about this parameter; because such a prior will usually overestimate mean branch lengths, it was predicted that flat priors would produce excessively high posterior probabilities on trees [132]. Whether flat branch length priors actually produce high posterior probabilities has, however, not been tested. Finally, previous studies have considered only a single pattern of branch lengths; different branch length patterns might interact with prior assumptions to produce different effects.

Chapters 5 and 6 examine the robustness of BMCMC to assumption violations likely to be encountered in the analysis of empirical sequence data. Chapter 5 specifically addresses the recent concerns that not sampling unresolved trees can lead to an unreliable assessment of statistical support using Bayesian techniques [67, 132]. I show that these concerns are in fact unfounded; not explicitly sampling unresolved trees does not undermine the reliability of existing BMCMC methods. Chapter 6 examines the effects of prior assumptions on posterior probabilities. I show that prior uncertainty can affect posterior probabilities over a range of evolutionary conditions. Specifically, if branch lengths are not known in advance, posterior probabilities calculated using a variety of prior assumptions can deviate strongly from those that would be inferred given perfect prior knowledge. The pattern of branch lengths on the true tree determines both the magnitude and direction of this effect, with some patterns skewing posterior probabilities downward and others skewing posterior probabilities upward. I also show that an empirical Bayes approach that fixes branch lengths at their maximum likelihood values can produce more reliable results than traditional Bayesian techniques that integrate over a variety of branch length values. Little is known about the reliability of phylogenetic confidence estimators, let alone their robustness to assumption violations. My results suggest that further research in this area is clearly warranted



if we wish to accurately gauge the confidence we should have in reported phylogenies.

## CHAPTER II

# PERFORMANCE OF MAXIMUM PARSIMONY AND LIKELIHOOD PHYLOGENETICS WHEN EVOLUTION IS HETEROGENEOUS

This chapter was originally published in the journal *Nature* (vol. 431, pp. 980-984, 2004). It was co-authored by Joseph W. Thornton, who assisted with experimental design and edited the manuscript.

### 2.1 Introduction

All inferences in comparative biology depend on accurate estimates of evolutionary relationships. Recent phylogenetic analyses have turned away from maximum parsimony towards the probabilistic techniques of maximum likelihood and Bayesian Markov Chain Monte Carlo (BMCMC). These probabilistic techniques represent a parametric approach to statistical phylogenetics, because their criterion for evaluating a topology—the probability of the data, given the tree—is calculated with reference to an explicit evolutionary model from which the data are assumed to be identically distributed. Maximum parsimony can be considered nonparametric, because trees are evaluated on the basis of a general metric—the minimum number of character state changes required to generate the data on a given tree—without assuming a specific distribution [98]. The shift to parametric methods was spurred, in large part, by studies showing that although

both approaches perform well most of the time [45], maximum parsimony is strongly biased towards recovering an incorrect tree under certain combinations of branch lengths, whereas maximum likelihood is not [21, 32, 47]. All these evaluations simulated sequences by a largely homogeneous evolutionary process in which data are identically distributed, and the correct evolutionary model was used to analyze data using maximum likelihood. There is ample evidence, however, that real-world gene sequences evolve heterogeneously and are not identically distributed.

Functional constraints on sites in a gene sequence often change through time, causing shifts in site-specific evolutionary rates, a phenomenon called heterotachy (meaning ‘different speeds’) [27, 48, 55, 71, 72, 75, 76, 84, 87, 91]. Current models available for phylogenetic inference assume the evolutionary process to be highly homogeneous across sites and so do not incorporate heterotachy. When a largely homogeneous evolutionary framework is imposed on sequences that evolve heterogeneously, parameter estimates are compromises over sites and lineages and are therefore incorrect for many or all sites. Likelihood-based techniques are guaranteed to recover the true phylogeny only when the correct model is used, and nonparametric statistical methods are often applied when the assumptions of parametric techniques are violated. On the other hand, parametric methods, including maximum likelihood, are generally more powerful than nonparametric techniques and can be robust to certain violations [16, 109].

In this chapter we show that maximum likelihood and BMCMC can become strongly biased and statistically inconsistent when the rates at which sequence sites evolve change non-identically over time. Maximum parsimony performs substantially better than current parametric methods over a wide range of conditions tested, including moderate heterogeneity and phylogenetic problems not normally considered difficult.

## 2.2 Methods

To determine the potential effects of heterogeneous evolution on phylogenetic accuracy, we simulated molecular sequence data in which different sites in the

sequence evolve under different evolutionary dynamics. Replicate data sets were analyzed using the model-based techniques maximum likelihood (ML) and Bayesian Markov Chain Monte Carlo (BMCMC) as well as the nonparametric method maximum parsimony (MP). To determine the accuracy of each method examined, we calculated the proportion of data sets from which the correct tree was uniquely recovered as the amount of phylogenetic signal (internal branch length) was systematically increased. A method that uniquely identifies the correct tree with high probability given a short internal branch length is considered more accurate than a method that requires more phylogenetic signal in order to identify the correct phylogeny. To identify the specific effects of heterogeneity on phylogenetic accuracy, results from heterogeneous sequences were compared to homogeneous controls simulated under similar conditions but without among-site heterogeneity.

### 2.2.1 Simulations

We simulated sequences along a 4-taxon tree  $((A,B),(C,D))$  (Figure 2.1a) with two independent partitions that were concatenated into one heterogeneous alignment. In one partition, long terminal branches ( $p \in [0.3, 0.75]$ ) lead to A and C, and short terminals ( $q \in [0.001, 0.4]$ ) lead to B and D. In the other partition, terminal branches to B and D have length  $p$ , whereas A and C have length  $q$ . The internal branch length ( $r \in [0.0, 0.5]$ ) is equal in both partitions. The two partitions were of equal size unless otherwise noted. Two-hundred replicate alignments of 1,000, 5,000, 10,000 and 100,000 characters were simulated under each set of conditions using the JC69 (DNA) or Poisson (protein) model. Average homogeneous control data were simulated using the same internal branch length as in the experimental condition and terminals with the mean length over the two partitions. Single-partition homogeneous controls were simulated using conditions for one of the experimental partitions (Figure 2.1a). Sequences were also simulated on 8-taxon trees derived from 4-taxon trees by bisecting each terminal branch at the halfway point.

### 2.2.2 Phylogenetic Analyses

Phylogenies were analysed using maximum parsimony (MP, provided by PAUP\* v4.0b10 [98]), maximum likelihood (ML, implemented by PAML v3.14 [130]) and Bayesian Markov Chain Monte Carlo (BMCMC, implemented by MrBayes v3.0b4 [95]). ML phylogenetic analysis was conducted with exhaustive topology searches, using a branch length smoothing delta cutoff of  $10^{-50}$  and a maximum of 1000 smoothing passes. Starting parameter values were randomly chosen, and branches were collapsed to zero length if optimized to  $\leq 10^{-8}$ . BMCMC analyses were conducted using 4 chains (*temp* = 0.2). To evaluate the stationarity of BMCMC chains, 100 randomly-selected heterogeneous datasets were individually examined. For each dataset, the minimum generation whose log-likelihood score was  $\geq$  the average log-likelihood score of the last 50 Markov chain samples was recorded as the generation in which the chain reached stationarity; after stationarity, the chain presumably samples from the correct posterior distribution. Based on this analysis, the first 5000 generations—a point always well past stationarity—were discarded as burnin to prevent starting conditions from affecting resulting phylogenetic estimates. Chains were run for 55,000 generations, sampled every 100 generations, and chain swapping was attempted every generation. Longer BMCMC runs of 105,000 generations were also assayed, but the resulting estimates of parameter means and variances did not change. The branch length prior probability distribution was assumed to be uniform on [0, 10], and topology prior probabilities were assumed to be equal.

We selected best-fit probabilistic models for ML and BMCMC analyses using a hierarchical likelihood ratio test on a random sample of 100 experimental datasets (using Modeltest v3.06 [89],  $\alpha = 0.05$ ). This analysis supported the true JC69 model as the best-fit model in 93% of tests (5% K80, 2% F81), so this model was used in all nucleotide analyses. The gamma and invariant sites models provided no increase in likelihood and were rejected; we conducted analyses with and without these models, however, to determine whether they improved performance of likelihood-based methods. For the invariant sites model, the proportion of invariant

sites was estimated for ML and integrated over in BMCMC, with prior probability distribution uniform on  $[0, 1]$ . For gamma-distributed ASRV, a five-category discrete approximation was used, with shape parameter ( $\alpha$ ) estimated by ML and prior probability uniform on  $[0.5, 50]$  for BMCMC. The covarion model implemented in MrBayes was also used, with prior probabilities for covarion switch rates uniformly distributed on  $[0, 100]$ . The true Poisson model was used for protein analysis, and maximum parsimony used equal weights.

To determine support, we used nonparametric bootstrapping (1,000 replicates) for maximum parsimony and maximum likelihood and posterior probability for BMCMC, with a support cutoff value of 95% to construct strongly supported consensus trees.

### 2.2.3 Accuracy

The accuracy of each method was calculated as the proportion of replicates for which the correct topology was uniquely recovered ( $\phi$ ). Nonlinear regression was performed using the logistic equation  $\phi = 1/1+\exp((BL_{50} - r)H)$ , in which  $BL_{50}$  is the estimated internal branch length that produces 50% correct recovery, and  $H$  estimates the steepness of the performance curve. The significance of differences among  $BL_{50}$ s was examined by a *t*-test.

### 2.2.4 Bias and Error

The type I error rate for each method was determined by analysing data sets generated under strong heterotachy with zero-length internal branches and determining the fraction of replicates falsely resolved with 95% bootstrap or posterior probability support [115]. The presence of bias was determined by calculating the proportion of erroneous estimates consistent with each possible incorrect topology over all internal branch lengths. The intensity of bias was investigated by calculating the proportion of erroneous topology estimates consistent with each possible incorrect topology when a 95% support cutoff was imposed.

To determine the impact of homogeneous optimization of branch lengths on maximum likelihood error, we compared the standard maximum likelihood algorithm that estimates a single set of branch lengths ( $ML_{\text{homo}}$ ) with several partitioned maximum likelihood models with constrained branch lengths.  $ML_{\text{true}}$  constrains all branch lengths for each site to the true values used to simulate data sets.  $ML_{\text{term}}$  constrains the internal branch lengths to the true value for each site, but terminal branches have the lengths homogeneously optimized under  $ML_{\text{homo}}$ .  $ML_{\text{short}}$  assumes the true internal and long terminal branches but uses the short terminal length from  $ML_{\text{homo}}$ .  $ML_{\text{long}}$  constrains the internal and short terminal branches to their true values and takes the long terminal branch lengths from  $ML_{\text{homo}}$ .

Support for the true topology by each character-state pattern was calculated from a 100,000-site data set constructed under strong heterotachy ( $p = 0.75, q = 0.05, r = 0.254$ ). Net support for the true tree is defined as the likelihood ratio of the true tree to the incorrect tree for each pattern  $x$ , weighted by the frequency of  $x$  ( $f(x)$ ) in the data set:

$$S_{((AB),(CD)),x} = \frac{P(x|((AB),(CD)))}{P(x|((AC),(BD)))} f(x).$$

To determine the performance impact of violating the identical distribution assumption when the true evolutionary model is used, we implemented a novel likelihood model ( $BMC_{\text{hetero}}$ ) that incorporates heterotachy a posteriori by applying two sets of branch lengths to the data. The  $BMC_{\text{hetero}}$  method was implemented by modifying the source code of MrBayes v3.0b4 to optimize and calculate a likelihood score conditioned on two independent sets of branch lengths for the same tree topology. For each sequence site  $x_i$  the likelihood of tree  $t$  with branch length sets  $b_1$  and  $b_2$  is

$$L(t|x_i) = \sum_{j=1}^2 \rho_{i,j} \times P(x_i|t, b_j),$$

where  $\rho_{i,j}$ —the posterior probability that  $x_i$  is in branch length set  $b_j$ —is calculated

from the data as

$$\rho_{i,j} = \frac{P(x_i|t, b_j)}{\sum_{k=1}^2 P(x_i|t, b_k)}.$$

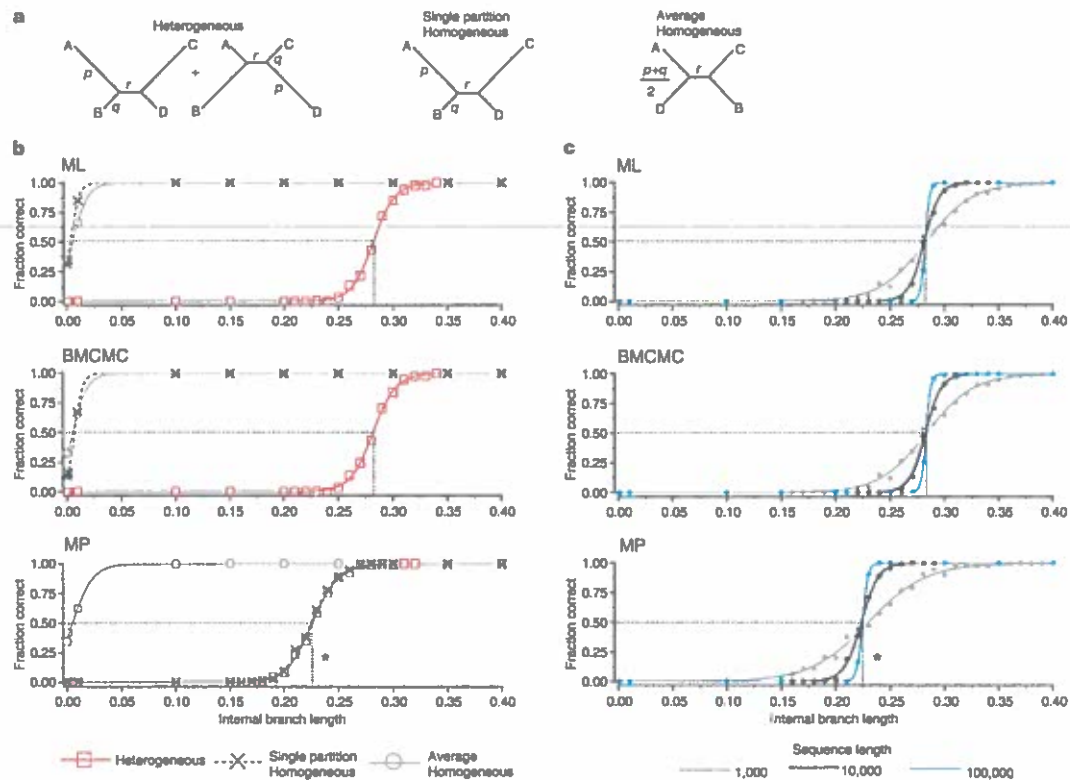
Branch length sets are proposed and accepted or rejected using the BMCMC algorithms already implemented in MrBayes, as is calculation of the overall posterior probability of each topology. Because of the increased number of parameters in this model, BMCMC<sub>hetero</sub> analyses were conducted using 8 chains to avoid local optima. All other parameter settings were equivalent to those we used to run standard BMCMC analyses under the JC69 model.

The BMCMC<sub>true</sub> method uses a priori partitioning and does not assume an identical distribution. Data are partitioned into two mutually-exclusive subsets corresponding to the partitions in which the data were simulated. For each partition, a single set of branch lengths is proposed at each generation, and the likelihood at each site is calculated assuming those lengths using the standard homogeneous algorithm. The total likelihood of the tree given the two partitions is the product of the likelihood over all sites in both partitions [129]. Branch lengths are proposed and accepted using the existing BMCMC techniques in MrBayes, and all settings were as described above for BMCMC<sub>hetero</sub>.

## 2.3 Results

We used an experimental approach to evaluate the phylogenetic accuracy of parametric and nonparametric methods under a simple form of heterotachy. We simulated replicate DNA sequence alignments with two symmetrical rate partitions along a four-taxon tree; each partition represents a phylogenetically challenging problem—two clades, each consisting of a long branch (length  $p$ ) and a short branch (length  $q$ )—but the sites with accelerated rates differ between partitions (Figure 2.1a). To reveal the specific impact of heterogeneity, we compared phylogenetic accuracy (the fraction of replicates from which the true tree was recovered) on heterogeneous data with accuracy on control sequences simulated under corresponding evolutionary conditions without heterogeneity (see Methods).

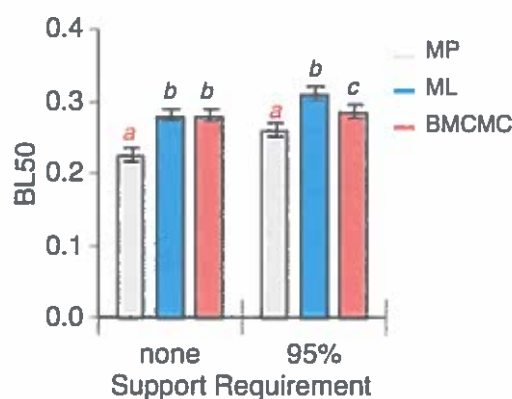




**FIGURE 2.1:** Likelihood-based methods are less accurate than maximum parsimony (MP) under heterogeneous conditions. **a**, Trees on which heterogeneous and control sequences were simulated. **b**, Heterotachy reduces the accuracy of likelihood methods. Accuracy is plotted against internal branch length for sequences with and without strong heterotachy. Dotted lines, BL<sub>50</sub> for each method (asterisk: maximum parsimony < maximum likelihood (ML) and BMCMC,  $P < 0.001$ ). **c**, Likelihood methods are inconsistent below the BL<sub>50</sub> under strong heterotachy, recovering the incorrect tree with increasing frequency as the amount of data increases.

Under conditions of strong heterotachy ( $p = 0.75$  substitutions per site,  $q = 0.05$ ), the accuracy of both maximum likelihood and BMCMC is dramatically reduced compared with homogeneous controls (Figure 2.1b). Both methods have zero accuracy when the internal branch length  $r < 0.22$ , and they reach 100% accuracy only when  $r > 0.34$ . Maximum parsimony is superior to the parametric methods when  $0.15 < r < 0.35$ , and it is never inferior. For each method, we used nonlinear regression to estimate the internal branch length at which 50% accuracy is achieved (BL<sub>50</sub>) and found that maximum parsimony can reliably recover the true

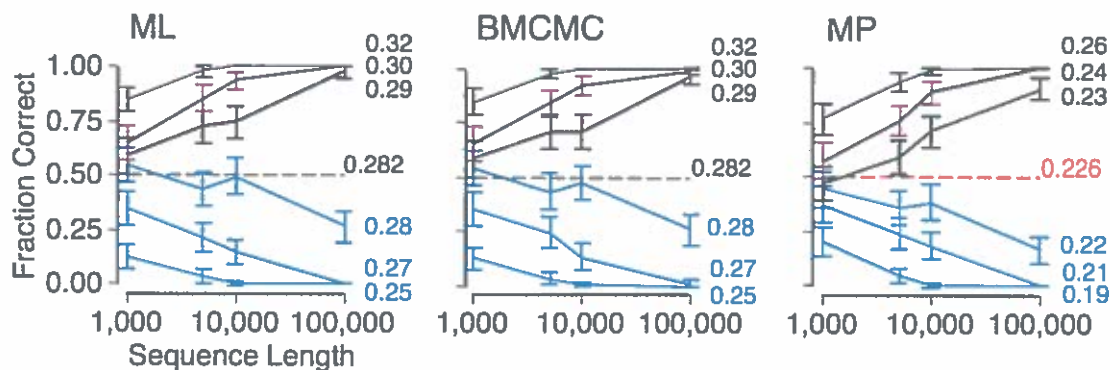
topology at significantly shorter internal branch lengths ( $BL_{50} = 0.22$ ) than the two likelihood-based methods ( $BL_{50} = 0.28$ ,  $P < 0.001$ ). Maximum parsimony's performance is worse than that of the parametric methods on single-partition data (due to the well-known long branch attraction bias [47]), but it is not additionally hampered by evolutionary heterogeneity ( $P = 0.76$ ). Maximum parsimony retains its performance advantage over maximum likelihood and BMCMC on heterotachous data when strong support is required to accept a tree as resolved (bootstrap or posterior probability  $> 95\%$ , Figure 2.2). These results indicate that heterotachy substantially reduces the accuracy of maximum likelihood and BMCMC on phylogenetic problems that are not difficult enough to impair maximum parsimony.



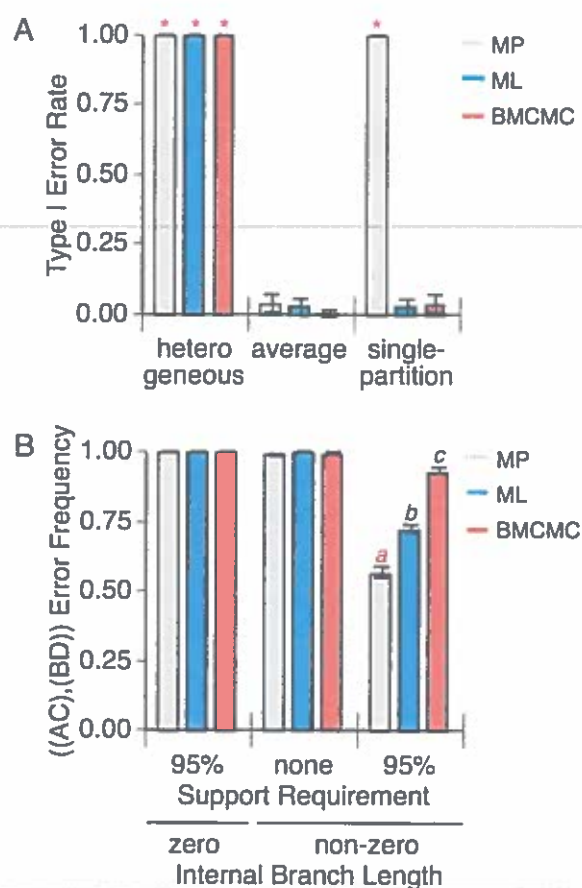
**FIGURE 2.2:** Maximum parsimony outperforms likelihood-based methods when strong support is required. Each method's  $BL_{50}$ —the internal branch length at which the true tree is recovered from half of replicates with the required support level—is shown under conditions of strong heterotachy (see Methods). In the 95% support category, trees were considered resolved only when supported by  $> 95\%$  nonparametric bootstrap proportions for MP and ML or by  $> 95\%$  posterior probability for BMCMC. Within each support category, significantly different results are indicated by different letters ( $P < 0.001$ ). 99% confidence intervals are shown. Unequal  $BL_{50}$ s for ML and BMCMC is likely caused by differences between nonparametric bootstrapping and posterior probabilities.

Under the heterotachous conditions studied, maximum likelihood and BMCMC are statistically inconsistent, converging on the wrong answer as the amount of data grows. For internal branch lengths below the  $BL_{50}$ , accuracy declines to zero as sequence length increases, indicating that parametric methods are statistically inconsistent in this region of parameter space; the  $BL_{50}$  therefore represents an inconsistency point (Figure 2.1c and Figure 2.3). This inconsistency is due to a directional bias: maximum likelihood and BMCMC specifically infer the erroneous tree ((AC),(BD)) with high support when the internal branch is shorter than the  $BL_{50}$ , including length zero (Figure 2.4). This is the same tree towards which maximum parsimony is biased on single-partition data, but heterotachy causes likelihood-based methods to infer the incorrect tree over a wider range of parameter values and with stronger apparent support.

Heterotachy reduces the performance of parametric methods across a broad range of evolutionary conditions. Whenever the short terminal branch length  $q < 0.3$ , maximum parsimony significantly outperforms both likelihood-based methods. Even fairly weak heterotachy—a ratio of branch lengths among partitions



**FIGURE 2.3:** The  $BL_{50}$  defines a method's inconsistency point. At internal branch lengths below the  $BL_{50}$ , increasing sequence length reduces accuracy, while increasing sequence length improves accuracy at internal branch lengths above the  $BL_{50}$  ( $BL_{50}$  indicated by dotted line, internal branch lengths for each series shown at right). Bars indicate 99% confidence intervals.



**FIGURE 2.4:** Evolutionary heterogeneity biases likelihood-based methods. **A)** Heterogeneity increases the type I error rate of likelihood-based methods. The proportion of zero-internal-branch-length datasets falsely resolved with > 95% support is shown for each method and data type. Error rates significantly > 0.05 ( $P < 0.001$ ) are indicated by asterisks. **B)** Erroneous inferences are biased towards a specific topology. The proportion of ((AC),(BD)) inferences among erroneous trees is shown for both zero- and non-zero- length internal branch datasets when strong nodal support is and is not required. Significantly different values ( $P < 0.001$ ) are indicated by different letters. Support is measured by nonparametric bootstrapping for MP and ML and by posterior probability for BMCMC. Bars indicate 99% confidence intervals.

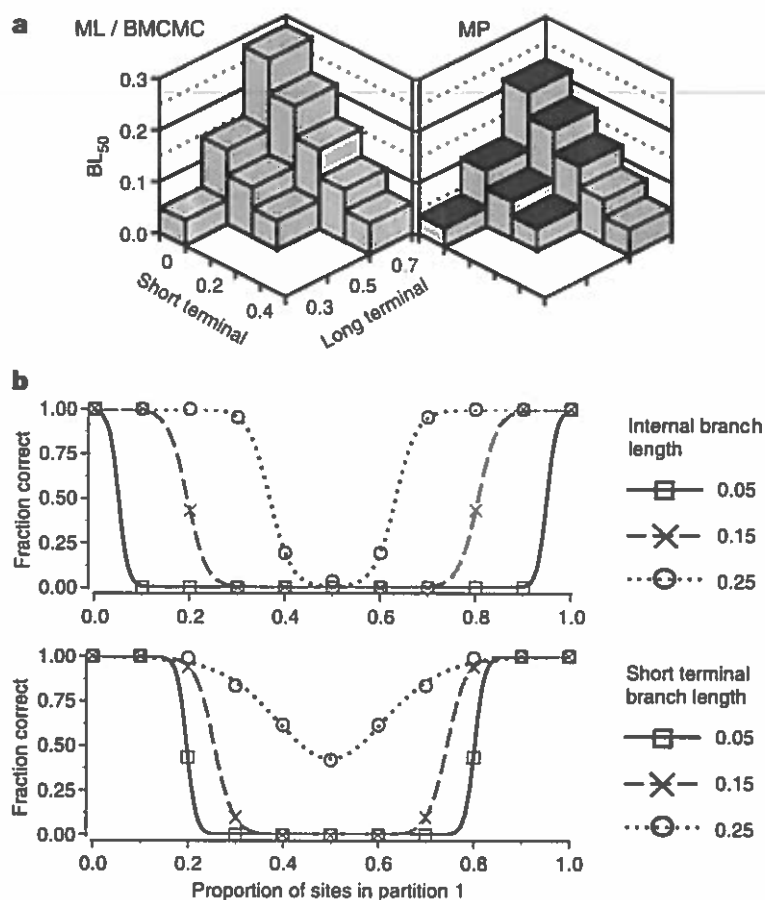
as low as 0.5:0.2—is sufficient to produce a significant performance disparity between the likelihood-based methods and maximum parsimony (Figure 2.5a). The more intense the heterotachy, the greater the performance difference. Furthermore, maximum likelihood’s accuracy can be reduced to zero when only a small fraction of

sites deviate in rate from the rest of the sequence. Fewer heterotachous sites are required to impair performance as heterotachy grows more intense or the phylogenetic problem becomes more difficult (Figure 2.5b).

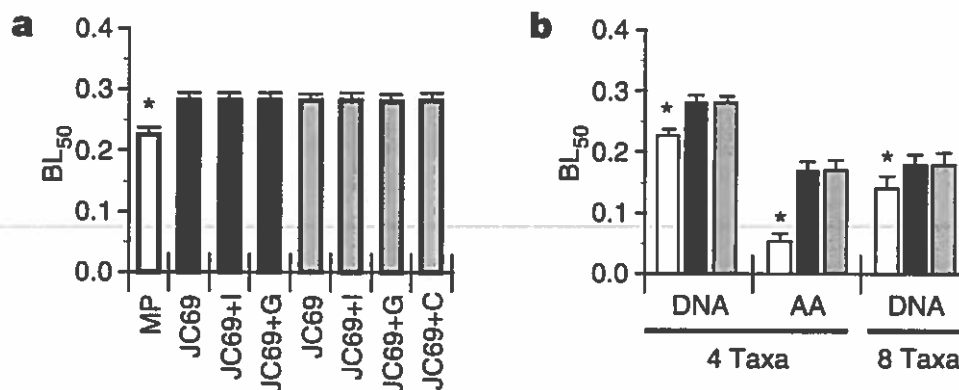
We used several existing likelihood models that account for among-site or among-lineage rate variation by applying identically distributed models of heterogeneity, including gamma, invariant sites and covarion models, but none improve the performance of maximum likelihood or BMCMC (Figure 2.6a). Using amino acid instead of nucleotide sequences substantially increases the accuracy of maximum parsimony ( $BL_{50} = 0.08$ ) because convergence is less likely with 20 than with 4 possible states. In contrast, maximum likelihood and BMCMC improve to a much smaller extent ( $BL_{50} = 0.18$ ). As a result, maximum parsimony's performance advantage on heterotachous protein sequences is even greater than on DNA (Figure 2.6b). Denser taxon sampling to break up long branches [43] improves the accuracy of all methods by about equal proportions (Figure 2.6b).

The accuracy of likelihood-based methods declines because they erroneously impose homogeneous branch lengths across sites. On heterotachous data with internal branch lengths below the inconsistency point, maximum likelihood underestimates the length of the internal branch on the correct tree and infers the lengths of the long and short terminals as approximately the average over the two partitions (Figure 2.7). To test whether these errors are responsible for phylogenetic bias, we compared the standard homogeneous maximum likelihood model ( $ML_{\text{homo}}$ ) with an a priori partitioned model in which the branch lengths for each site are constrained to their true values ( $ML_{\text{true}}$ ). As Figure 2.8a shows,  $ML_{\text{true}}$  has much better performance ( $P < 0.001$ ). Models that set only the internal ( $ML_{\text{term}}$ ) or the internal and long terminal branches ( $ML_{\text{short}}$ ) to their true lengths did not improve performance. Correcting the short terminal ( $ML_{\text{long}}$ ), however, yields a substantial improvement in phylogenetic accuracy. Erroneous optimization of the short terminal length using 'compromise' branch lengths is therefore the primary cause of heterotachy-induced phylogenetic error in maximum likelihood (Figure 2.8a).

Maximum likelihood's bias is caused by misinterpretation of specific character state patterns. We analysed the contribution each character state pattern makes to



**FIGURE 2.5:** Parsimony outperforms likelihood over a wide range of heterotachous conditions. **a**, Maximum parsimony is more accurate than likelihood-based methods on data with weaker heterotachy. Bars show the BL<sub>50</sub> for combinations of long and short terminal branch lengths in heterotachous data sets (black: maximum parsimony < maximum likelihood and BMCMC,  $P < 0.001$ ). The BL<sub>50</sub>s for maximum likelihood and BMCMC are equivalent for all conditions ( $P > 0.91$ ). **b**, Maximum likelihood accuracy is impaired when only a small fraction of sites are heterotachous. Accuracy is plotted against the fraction of heterotachous sites as the phylogenetic problem becomes more difficult (upper panel:  $p = 0.75, q = 0.05$ ) and heterotachy more intense (lower panel:  $p = 0.75, r = 0.15$ ).

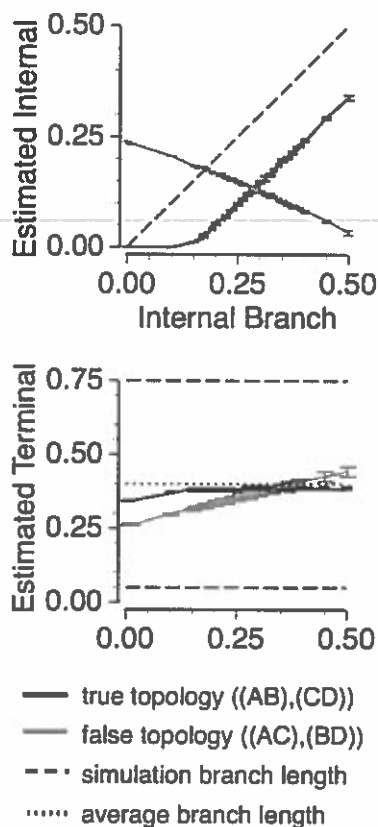


**FIGURE 2.6:** Maximum parsimony is more accurate than likelihood methods when techniques to improve phylogenetic performance are used. **a**, Accuracy of likelihood-based methods on heterotachous data does not improve when evolutionary models that incorporate among-site rate variation (+G, gamma distribution; +I, invariant sites) or covarion heterotachy (+C) are used. BL<sub>50</sub>s are shown under strong heterotachy; bars indicate 99% confidence intervals. Asterisks show lower BL<sub>50</sub> values ( $P < 0.001$ ). **b**, Maximum parsimony (white) outperforms maximum likelihood (black) and BMCMC (grey) on amino acid sequences and 8-taxon data sets with strong heterotachy.

the likelihood of the true and erroneous trees and compared net support for the true tree using  $ML_{\text{homo}}$  to that using the heterogeneous model  $ML_{\text{true}}$  (Figure 2.8b).

Patterns that provide the most support for the correct tree under  $ML_{\text{true}}$  ( $xyxy$  and  $xyyz$ ) only weakly support the true tree when  $ML_{\text{homo}}$  is used; this occurs because  $ML_{\text{homo}}$  overestimates the probability that these patterns are due to convergence on short-terminal branches whose lengths are overestimated. In contrast, the convergent patterns  $xyxy$  and  $xyxz$  support the wrong tree using either method. As a result, the likelihood of the incorrect tree becomes greater than that of the true tree when  $ML_{\text{homo}}$  is used on heterotachous data. Under the same conditions, maximum parsimony recovers the true tree because the frequency of  $xyxy$  is greater than that of  $xyxz$ ; the patterns  $xyyz$  and  $xyxz$ , which taken together mislead  $ML_{\text{homo}}$ , are not informative in a nonparametric context.

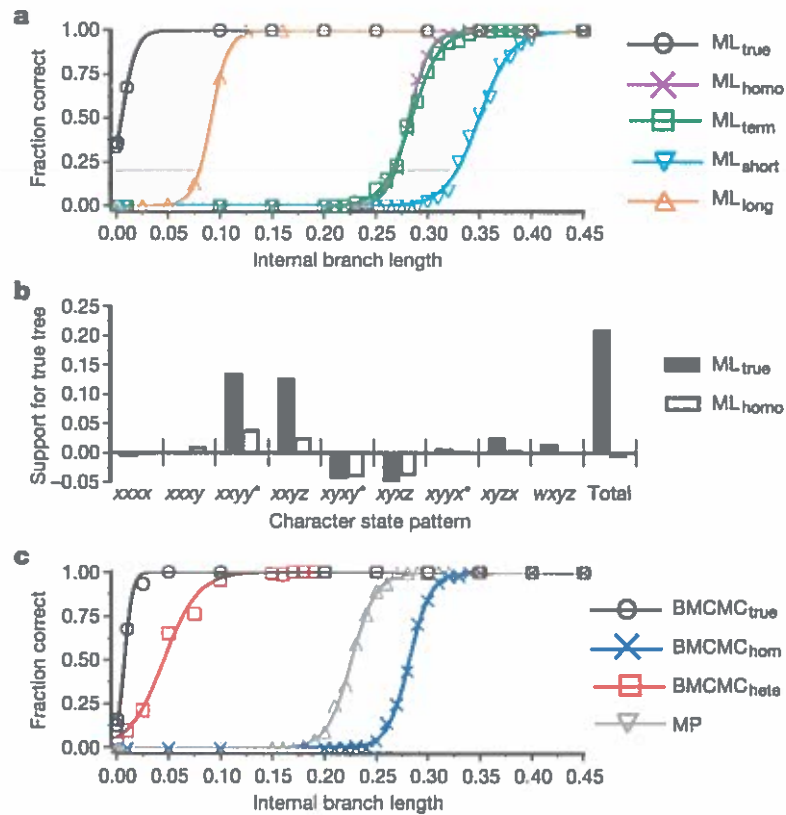
The bias of parametric methods arises due to heterogeneity in the data and the resulting violation of the identical distribution assumption, as predicted



**FIGURE 2.7:** Maximum likelihood fails to correctly infer the lengths of internal and terminal branches from heterotachous data ( $p = 0.75, q = 0.05, s = 0.50$ ). Top, ML-estimated internal branch lengths on the correct  $((AB),(CD))$  and incorrect  $((AC),(BD))$  topologies are compared to the true branch length (dashed line) as the simulated internal branch increases. Bottom, ML-estimated terminal branch lengths on the two topologies are shown as internal branch length increases. The true terminal branch lengths are indicated by dashed lines, while their average length is indicated by a dotted line. 99% confidence intervals are shown. Note that the lengths estimated on the correct and incorrect phylogenies intersect at the BL50.

theoretically [11]. We implemented a novel likelihood method using a mixed model (BMCMC<sub>hetero</sub>) that incorporates heterotachy by including two branch length sets for each topology. For each sequence site, the likelihood is calculated for each branch length set, weighted by the posterior probability of the site being in that set and then summed to yield the total likelihood. This model, which corresponds to the true evolutionary conditions but assuming an identical data distribution,



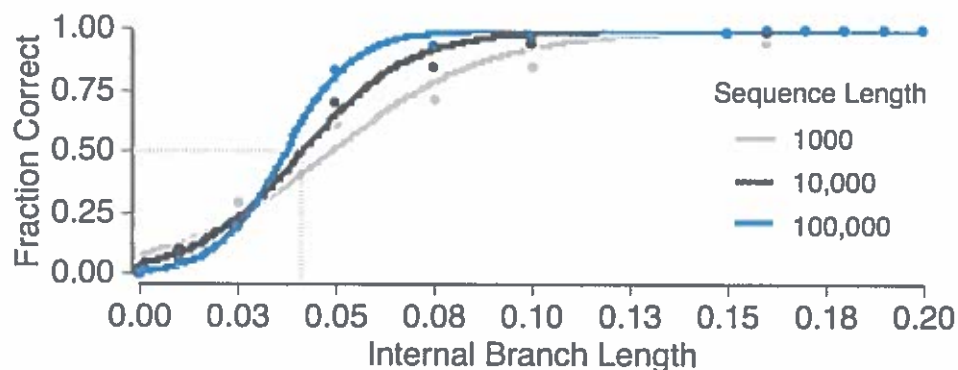


**FIGURE 2.8:** Poor maximum likelihood performance is due to assuming homogeneous branch lengths. **a**, Maximum likelihood error is caused primarily by overestimating short terminal branch lengths due to heterogeneity. Accuracy on strongly heterotachous sequences is shown as the internal branch length increases, using several likelihood models that constrain all ( $ML_{true}$ ), some ( $ML_{term}$ ,  $ML_{short}$ ,  $ML_{long}$ ) or no branches on the tree ( $ML_{homo}$ ) to their true lengths for all sites. **b**, Support for the true tree by specific character state patterns is reduced due to strong heterogeneity when  $ML_{homo}$  is used. For each character state pattern and model, net support is shown as the ratio of the likelihood of the true topology to the likelihood of the incorrect ((AC),(BD)) tree, weighted by the frequency of the pattern. Asterisks indicate parsimony-informative patterns. **c**, Incorporating heterotachy improves the accuracy of parametric methods. Accuracy on strongly heterotachous data are shown for the homogeneous model ( $BMCMC_{homo}$ ), a model that allows two independent branch length sets and correct a priori partitioning of sites ( $BMCMC_{true}$ ), and a novel model with two branch length sets and likelihoods calculated on the basis of a posteriori weighting ( $BMCMC_{hetero}$ ).

performs dramatically better than both maximum parsimony and the standard maximum likelihood or BMCMC algorithms ( $BL_{50} = 0.045$ ,  $P < 0.001$ ) on heterotachous data (Figure 2.8c). It did not perform as well, however, as a non-identically distributed method ( $BMCMC_{true}$ ) that uses the true evolutionary model with a priori sorting of sites into their true partitions. Furthermore,  $BMCMC_{hetero}$  remains statistically inconsistent, converging on the wrong tree as sequence length increases at internal branch lengths  $r < BL_{50}$ , (Figure 2.9).  $BMCMC_{true}$  is consistent under all conditions examined. These results indicate that violating the identical distribution assumption can cause inconsistency, even when a model approximating the ‘true’ evolutionary process is used.

## 2.4 Discussion

The form of heterotachy studied here is only one way that heterotachy can be distributed on a tree. Additional studies presented in chapter 3 indicate that several



**FIGURE 2.9:** Violating the identical distribution assumption causes likelihood-based methods to be statistically inconsistent when the correct heterotachous evolutionary model is used. Accuracy on strongly heterotachous data using a novel BMCMC method that assumes the correct number of branch length partitions but calculates likelihoods under an identically-distributed model ( $BMCMC_{hetero}$ , see Methods) is plotted against internal branch length for short, medium, and long sequences. At internal branch lengths below the  $BL_{50}$ , increasing sequence length reduces accuracy.  $BL_{50}$  (indicated by a dotted line) is the same for all sequence lengths ( $P = 0.89$ ).

other forms of heterotachy can also impair the accuracy of parametric methods. The evolutionary model used in our simulations is a simplified one; the extent to which phylogenetic accuracy is impaired by the more complex evolutionary dynamics likely to affect real-world sequences is currently unknown. There are numerous sequence data sets from which parametric methods have failed to infer otherwise well corroborated phylogenies [14, 80, 94, 96], including one in which heterotachy has recently been implicated [55].

There are two ways to avert the negative effects of heterogeneity on parametric methods. One is to use maximum parsimony, which is not affected by heterotachy because it does not assume an identically distributed evolutionary process. The other is to develop more complex parametric models. Our results indicate that a new likelihood method using mixed branch length models may offer substantially improved accuracy on heterotachous sequences, but there are reasons for caution. The model that performed well in our tests matched the true evolutionary process, which we knew a priori. With real sequences, we do not know the true number of branch length partitions, so imposed models may regularly use either too many or too few branch length parameters. For many sequences, the actual number of branch length categories may approach the number of sites; under these conditions, the true one-category-per-site likelihood model is formally equivalent to maximum parsimony [119]. Finally, the computational burden of mixed-model phylogenetic inference grows exponentially with the number of branch length sets. With current algorithms and computing power, incorporating heterotachy into a likelihood framework will often require sacrifices in the number of sequences analysed or the rigor with which tree and parameter space are searched, which may also reduce phylogenetic accuracy [98]. The issues of determining how many branch length sets to use in a mixed model analysis and how to speed up mixed model computations are addressed in chapters 3 and 4.

Our findings place those who infer and use phylogenetic trees in an uncertain position. Previous research has shown that parametric methods are superior or equal to nonparametric approaches when evolutionary heterogeneity is not present, but our work shows that maximum parsimony can substantially outperform current

likelihood-based methods when it is. Worse still, heterotachy-induced bias leaves no obvious signature, because the inferred trees have moderate branch lengths and strong support for erroneous nodes. With no reliable a posteriori diagnostic for heterotachy-induced phylogenetic error, how can we know which method to choose or, when trees from different methods conflict, which one to favor? The overall frequency and severity of the conditions that favor likelihood as compared with those that favor parsimony is not yet known for real-world sequences. At present, we recommend reporting nonparametric analyses along with parametric results and interpreting likelihood-based inferences with the same caution now applied to maximum parsimony trees. In the future, it is possible that new mixed-model techniques may improve likelihood's performance to the point that it is consistently superior to nonparametric methods.

## CHAPTER III

# EFFECTS OF HETEROTACHY ON STANDARD AND MIXED BRANCH LENGTH PHYLOGENETIC INFERENCE

### 3.1 Introduction

The molecular evolutionary process is complex and dynamic; in fact, the more closely examined the process is, the more complex and dynamic it appears to be. While it has been expected for some time that different sites in a sequence may have evolved at different evolutionary rates, and that evolutionary rates may have changed over time independently for different sites in different lineages [26, 27], it is now well established that molecules commonly evolve under such “heterotachous” dynamics [5, 31, 48, 69, 71, 72, 75, 76, 77, 81, 84, 91]. Understanding the properties of complex evolutionary dynamics such as heterotachy is important not only for improving our understanding of how molecules have evolved but also because heterotachy can potentially limit the accuracy with which evolutionary inferences—phylogenetic trees and parameter values describing the evolutionary process—can be made using existing techniques [55, 62, 68, 82].

It is well known that inference techniques relying on a model of the evolutionary process can produce biased phylogenetic inferences when that model is not the same as the true process [49, 65, 115, 131]. Recent simulation studies [28, 30, 85, 104] have confirmed our initial findings that some forms of heterotachy can strongly

mislead existing model-based inference techniques that do not accommodate heterotachous evolution [62]. Heterotachy has also been implicated in the failure of existing techniques to recover the correct Microsporidia + Fungi grouping from elongation factor 1 $\alpha$  sequences [55] and may be an important factor contributing to other phylogenetic errors [9, 83]. Due to the prevalence of heterotachy in molecular evolution and its potentially negative effects on the accuracy of phylogenetic inference, understanding the causes of phylogenetic errors when heterotachy is not incorporated, and developing techniques that overcome these problems, is a major concern.

Using a controlled set of simulation experiments, we have shown that misestimation of site-specific branch lengths—particularly overestimation of short terminal lengths—caused by one type of heterotachy can cause model-based methods (maximum likelihood and Bayesian techniques) to misinterpret phylogenetic signal as convergent evolution, resulting in a strong bias in favor of an incorrect tree [62]. Other studies have largely focused on elucidating the relative accuracy of model-based inference and maximum parsimony (MP) under different types of heterotachy [28, 30, 69, 85, 104]—the consensus being that model-based techniques are more accurate than MP more often than the converse—rather than determining the causes of phylogenetic errors and the effects of heterotachy on parameter estimation. As a result, although the rates of erroneous inferences under various simulation conditions have been well documented, the types of errors made under different conditions, as well as the reasons for these errors, remain largely unknown. Knowledge of the types of phylogenetic errors likely to occur as well as the specific evolutionary conditions responsible for such errors is required in order to diagnose potentially incorrect inferences made from real data and guide the development of new models more robust to heterotachy.

As part of our initial study, we introduced a general model for incorporating heterotachy by allowing each branch on the tree to have multiple lengths; we noted that this model produced significantly more accurate phylogenies than either standard evolutionary models or maximum parsimony under the conditions we examined [62]. Spencer et al. [104] subsequently improved on our initial design and

showed that this mixed branch length model should be statistically consistent, converging on the correct tree with probability 1.0 as sequence length approaches infinity, provided the model does not contain too many parameters to uniquely identify the correct tree. Although a promising approach for incorporating heterotachy, the use of highly complex evolutionary models such as the mixed branch length model raises a number of potential concerns [62, 107, 118]. First, when the correct number of branch length sets is not known in advance, the number of lengths per branch must be estimated from the data using statistical techniques such as Likelihood Ratio Test (LRT), Akaike Information Criterion (AIC), or Bayesian Information Criterion (BIC); the accuracy with which these tests estimate the correct number of branch lengths is currently unknown. Second, as the number of parameters in the model grows, the ability to accurately estimate the values of these parameters from finite data decreases, leading to a potential loss of resolution or inability to uniquely identify the correct phylogeny. It is not known whether this is likely to be a problem for mixed branch length analyses of real data. Third, model-based techniques are inherently computationally expensive [98], and more complex models generally require even greater computer time in order to estimate the additional parameters. Whether mixed branch length models can be solved quickly enough to apply to real phylogenetic problems has not been determined. Finally, it has been suggested that heterotachy in real sequence data may not fit a mixed-branch-length model [69, 70]; instead, heterotachy may be more appropriately envisaged as lineage-specific changes in the proportion of invariant sites. The mixed branch length model has not been tested in such cases.

Here we conduct detailed simulation experiments to examine the causes of phylogenetic error observed for a number of types of heterotachy. We observe that, in general, various forms of heterotachy can negatively affect the accuracy of existing model-based phylogenetic approaches, leading to misestimation of both phylogeny and expected branch lengths. In addition, we develop a software implementation of the mixed branch length model and examine its accuracy under both simulated and real-world conditions. We show that although computation times are dramatically increased compared to simpler models, a mixed branch

length model can provide highly accurate phylogenetic inferences over a wide range of challenging heterotachous conditions.

## 3.2 Methods

Experiments and analyses conducted in this chapter follow the methodology described in chapter 2.

### 3.2.1 Phylogenetic Analysis

Sequence alignments were analyzed using maximum likelihood (ML), Bayesian Markov Chain Monte Carlo (BMCMC), and unweighted maximum parsimony (MP). MP analyses and ML analyses on nucleotide data were conducted using exhaustive topology search in PAUP\* 4.0b10 [114]. For ML analyses, the best-fit evolutionary model was selected by hierarchical likelihood ratio test implemented in Modeltest 3.7 [89], with  $\alpha=0.05$ . Bayesian analyses were conducted using MrBayes 3.1 [95]. Two independent runs of four chains were run until the average standard deviation in posterior probabilities dropped below 0.01; the first 5,000 generations were discarded as burnin to reduce the effects of starting conditions on posterior probabilities. Topology priors were equal for each resolved tree, branch length priors were uniform on [0,10], and the default priors were used for other model parameters.

In addition to standard ML, we analyzed heterotachous data using a mixed branch length model which calculates likelihoods using multiple independent sets of branch lengths on the tree [62, 104]. The likelihood of tree  $t$  given data  $X$  and branch length sets  $b=(b_1, b_2, \dots, b_n)$  is given by

$$L(t|X) = \sum_{i=1}^n \rho_i P(X|t, b_i)$$

where each  $\rho_i$  is estimated from the data, and  $P(X|t, b_i)$  is the likelihood of the tree given branch lengths  $b_i$ .



Mixed branch length model analyses were conducted in a maximum likelihood framework using a novel simulated annealing algorithm to optimize the tree topology and all model parameters. The simulated annealing algorithm is incredibly simple: given a set of model parameters at iteration  $i$ , the algorithm slightly alters some of the parameters to produce a new proposal ( $i + 1$ ). This proposal is either accepted or rejected based on the Metropolis criterion: proposals that improve the likelihood score are always accepted, while proposals that reduce the likelihood are accepted with probability proportional to the likelihood ratio between the new proposal and the old state. As the algorithm runs, a *temperature* parameter controls the probability of accepting proposals that reduce the likelihood score. At the beginning, ‘bad’ proposals are accepted with higher probability; this acceptance probability is gradually reduced until, by the end of the run, bad moves are rarely if ever accepted [60].

We used a temperature annealing schedule with a geometric descent of 1000 temperatures starting from 1.0 and ending at  $10^{-5}$ . At each temperature, 1000 parameter changes were attempted, with acceptance based on the Metropolis criterion; topology rearrangements included TBR, SPR, and NNI. The best-fit number of branch length categories ( $n$  above) was selected using AIC [1].

In addition to both standard and mixed-branch-length models, we performed phylogenetic analyses using the true maximum likelihood model ( $ML_{\text{true}}$ ), which correctly partitions sites into branch length categories and estimates branch lengths separately for each category. To validate the accuracy of our simulated annealing algorithm, we re-ran each mixed model analysis using the correct mixture proportions ( $\rho$  values) and branch lengths derived from analyses using the true evolutionary model ( $ML_{\text{true}}$ ); the results of these analyses were unchanged compared to simulated annealing estimation (not shown).

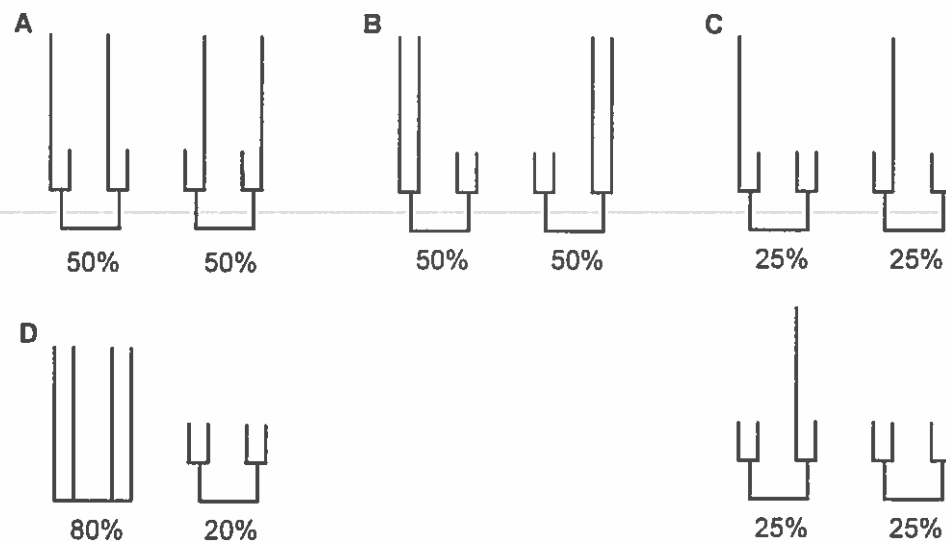
### 3.2.2 Branch Length Heterogeneity

We simulated 5,000-nucleotide datasets using the JC69 model under four types of four-taxon branch length heterogeneity (Figure 3.1): 1) Felsenstein Zone

Heterotachy (FZH), 2) Inverse-Felsenstein Zone Heterotachy (IFZH), 3) Single Long Branch Heterotachy (SLBH), and 4) Signal-Noise Heterotachy (SNH). Both FZH and IFZH partition sites into two branch length categories, with equal numbers of sites in each category. Long branches (0.75 substitutions/site) lead to two terminal lineages, while short branches (0.05) lead to the other two terminal lineages, but the lineages with long terminal branches are different in different branch length categories. In the case of FZH, the long terminal branches are not sister to one another, while long branches lead to sister taxa in IFZH. The internal branch length (which is the same in both categories) was varied between 0.0 and 0.4. SLBH consists of four branch length categories; in each category, a single lineage has a long terminal branch (0.75), while all other terminal branches are short (0.05). We varied both the internal branch length (0.0-0.4, the same in all branch length categories) and the proportion of sites in the first branch length category (0.2-1.0); the remaining sites were equally-proportioned among the other three branch length categories. SNH partitions sites into two categories; in the first category, sequences evolve with long terminal branch lengths (0.75) and a zero-length internal branch. In the other category, terminal branch lengths are short (0.05), and the internal branch length varies between 0.0 and 0.4. We also varied the proportion of sites in the first branch length category from 70% to 95% of sites. Two hundred replicate sequence alignments were simulated under each set of evolutionary conditions.

Phylogenetic analyses of simulated data were conducted as described above. For each phylogenetic method and set of simulation conditions, we calculated the proportion of replicates from which the correct phylogeny was uniquely recovered. We plotted the proportion uniquely correct for each method against increasing internal branch length and estimated the branch length at which 50% correct recovery was achieved (BL<sub>50</sub>) using nonlinear regression of the logistic equation:  $1/1+\exp((BL_{50} - r)H)$ , where  $r$  is the internal branch length and  $H$  estimates the steepness of the curve. We compared the accuracy of different methods by comparing BL<sub>50</sub> estimates; significance was assessed using a two-way  $t$ -test [62].

Bias was examined by simulating sequences under heterotachous conditions but with a zero-length internal branch. Five hundred replicate sequence alignments were



**FIGURE 3.1:** Summary of branch length heterogeneity simulations. Sequences of 5000 nt were simulated under each set of conditions (A-D) using the JC69 transition model. The proportion of sites evolving under each set of branch lengths is shown below the branch lengths used to simulate data. Long terminal branches were 0.75 substitutions/site, with short terminal branches of 0.05; the internal branch length was allowed to vary. Conditions were: A) Felsenstein Zone Heterotachy (FZH), B) Inverse-Felsenstein Zone Heterotachy (IFZH), C) Single Long Branch Heterotachy (SLBH), and C) Signal-Noise Heterotachy (SNH).

analyzed using each phylogenetic method, and we recorded the proportion of replicates from which each possible topology was recovered; an unbiased method should recover each possible tree with roughly equal proportions, and the significance of deviation from this behavior was assessed using a chi-square test. The severity of bias was assessed for standard ML, BMCMC, and MP by determining the proportion of replicates falsely resolved with support  $> 0.95$ . Support was assessed using nonparametric bootstrapping (1000 replicates) for ML and MP and posterior probabilities for BMCMC. In each case, deviation from an expected false-positive error rate of 0.05 was assessed using a one-sided  $t$ -test.

In order to assess the asymptotic performance of ML with infinite data, ideal pseudo-datasets with no stochastic error were analyzed. We calculated the expected

frequency of each character state pattern ( $f(x)$ ) under SLBH conditions with long terminal branch lengths of 0.75 substitutions/site, short terminals 0.05, and an internal branch length of 0.01. We implemented custom software to calculate likelihoods given this vector of state pattern frequencies as follows. The per-site likelihood of tree  $t$  given state pattern  $x$  is calculated by raising the probability of the pattern, given the tree, to the frequency with which that pattern is expected to occur:  $L(t|x) = P(x|t)^{f(x)}$ . The total per-site likelihood of the tree is the product of this partial likelihood over all possible state patterns. To determine the maximum likelihood estimate of the internal branch length when infinite data are available, we calculated the likelihoods of internal branch lengths between 0.0 and 0.01 substitutions/site, with other branch lengths optimized.

We also examined the accuracy with which different phylogenetic methods estimate expected branch lengths over sites from finite heterotachous data. For each set of simulation conditions, we calculated the expected internal and terminal branch lengths using standard ML, the mixed branch length model, and the correct ML model ( $ML_{true}$ ). For the mixed model and  $ML_{true}$ , expected branch lengths over sites were calculated by multiplying each site-specific branch length by the weight associated with that length; for the mixed model, these weights are inferred from the data, while weights are correctly assigned *a priori* for  $ML_{true}$ . In the case of terminal branches, we report the average branch length over all four terminals. Mean branch length estimates were calculated over 200 replicates for each set of simulation conditions.

### 3.2.3 Elongation Factor $1\alpha$ Sequences

We analyzed the Micro\* dataset of [55] using standard ML (JTT+gamma model), MP, BMCMC (JTT+gamma+covarian), and the mixed branch length model using JTT+gamma and a variable number of branch length categories (from 1 to 7). ML analyses were conducted using 4 gamma rate categories, with branch lengths and shape parameter optimized using the simulated annealing algorithm described above. We calculated the best-fit number of branch length categories for

the mixed branch length model using AIC and assuming either the artifactual Microsporidia + Archaeobacterial (MA) tree or the correct Microsporidia + Fungi (MF) topology. To determine the preferred tree for each number of branch length categories, we calculated the likelihood ratio MF/MA and assessed the support for the most likely hypothesis using the AU test [99] implemented in CONSEL [100].

Additionally, for the model selected by AIC, we calculated the posterior probability that each site evolved according to each set of inferred branch lengths. Assuming the correct MF tree ( $t$ ), the posterior probability of branch length set  $b_i$  for each site  $x$  ( $PP(b_i|t, x)$ ) was calculated by multiplying the proportion of sites expected to evolve under the given set of branch lengths ( $\rho_i$  above) by the likelihood obtained for that site under the inferred lengths ( $P(x|t, b_i)$ ) and dividing by the sum of the proportion-times-likelihood over all branch length sets:

$$PP(b_i|t, x) = \frac{\rho_i P(x|t, b_i)}{\sum_{j=1}^n \rho_j P(x|t, b_j)}$$

### 3.3 Results

In order to determine the potential effects of heterotachous evolution on phylogenetic inference, we simulated replicate datasets under a complex evolutionary process in which evolutionary rates are heterogeneously distributed across both sites and lineages. For each set of simulation conditions, we determined the frequency with which correct evolutionary inferences are made using widely available phylogenetic techniques: 1) standard maximum likelihood (ML) with the evolutionary model selected by likelihood ratio test, 2) Bayesian Markov Chain Monte Carlo (BMCMC) using the same model as for ML, and 3) equally-weighted maximum parsimony (MP). In addition to these conventional approaches, we examined the accuracy of two methods designed to address heterotachy: 1) a Bayesian implementation of the “covarion” model [120], which allows sites to switch between being variable (“on”) and invariant (“off”) as they evolve along the tree, and 2) a maximum likelihood implementation of a mixed branch length model [62, 104] allowing each branch on the tree to have multiple lengths; the likelihood

for each site is calculated as a weighted average over all lengths, with weights being estimated from the data (see Materials and Methods). The accuracy of these methods was compared to a control model ( $ML_{true}$ ) that precisely matches the simulation conditions by correctly determining site-specific evolutionary properties *a priori*.

### 3.3.1 Branch Length Heterogeneity

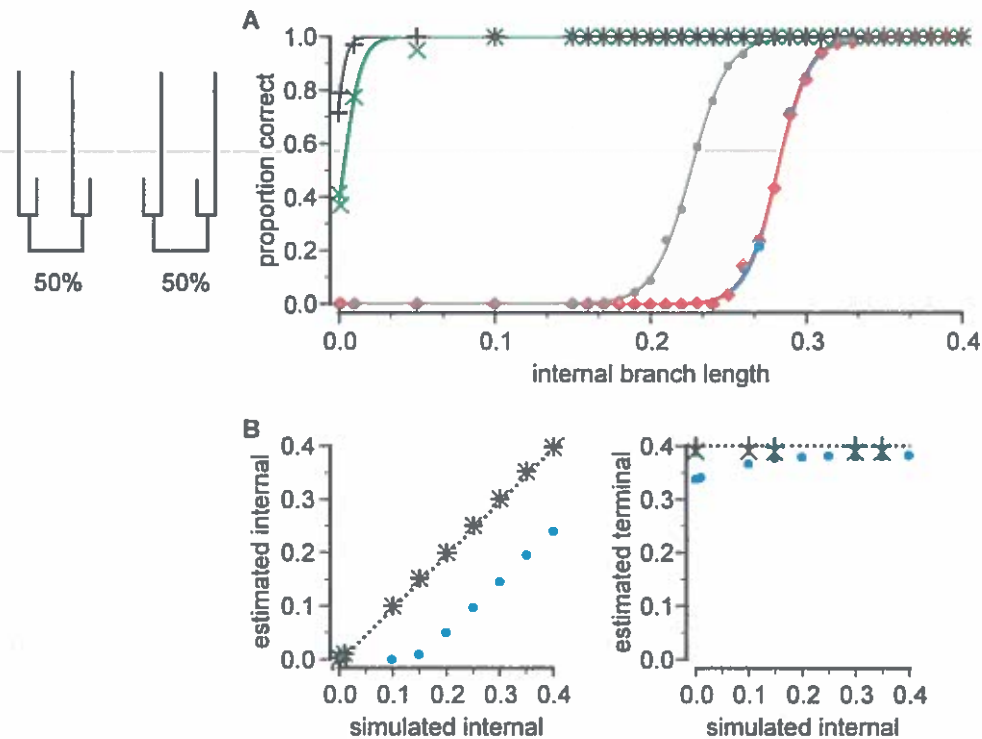
In its most general formulation, the concept of heterotachous evolution simply states that different sites evolve at different evolutionary rates, and the rate of each site can change independently as evolution proceeds [72, 84]. To simulate sequences under this general model, we allowed different sites in the alignment to evolve under completely independent branch lengths. Other ways of modeling heterotachy—such as a covarion process [26, 27, 120] or lineage-specific changes in the proportion of invariant sites [68, 69]—are special cases of this heterogeneous branch length model, so simulating sequences under heterogeneous branch lengths allows us to examine the potential effects of heterotachy without subscribing to a particular formulation of what real-world heterotachy might look like. Because different types of heterotachy (i.e., different combinations of heterogeneous branch lengths) might produce very different effects on the accuracy of phylogenetic methods, we examined datasets generated using a variety of branch length combinations.

#### Felsenstein Zone and Inverse-Felsenstein Zone Heterotachy (FZH and IFZH)

In both Felsenstein Zone Heterotachy (FZH) and Inverse-Felsenstein Zone Heterotachy (IFZH), sites are divided into two branch length classes, each of which contains two long terminal branches and two short terminal branches, but the lineages with accelerated rates differ between classes (see Figure 3.2 and Figure 3.3). The internal branch length is the same for all sites. The difference between FZH and IFZH is that long-branch lineages are sister to one another in IFZH, while they are not in FZH. Chapter 2 focused exclusively on elucidating the effects of FZH on

phylogenetic inference [62]; FZH conditions were found to strongly impair the accuracy of model-based techniques across a range of conditions. Here we simulated sequences of 5,000 nucleotides under strong FZH conditions (long terminal branch lengths 0.75 substitutions/site, short terminals 0.05, and a variable internal branch length) with equal proportions of sites in each branch length class. Data were analyzed using the phylogenetic techniques described above, and the internal branch length at which each method recovered the correct tree 50% of the time ( $BL_{50}$ ) was calculated using nonlinear regression (see Materials and Methods). We compared the accuracy of different methods to one another by comparing  $BL_{50}$ s, a lower  $BL_{50}$  indicates a more accurate phylogenetic method which can reliably infer the correct tree given less phylogenetic signal (i.e., a shorter internal branch).

Under FZH conditions, the true evolutionary model that correctly partitions sites into branch length classes *a priori* and estimates the branch lengths within each class separately ( $ML_{true}$ ) produced highly accurate phylogenies, with a  $BL_{50} < 0.001$  (Figure 3.2A). In contrast, the accuracy of both standard ML and BMCMC was severely reduced by FZH conditions ( $BL_{50} = 0.28$ ,  $P < 0.001$ ) when the model was selected using likelihood ratio test (LRT, which selected the correct JC69 model 93% of the time at  $\alpha = 0.05$ ). At internal branch lengths below the  $BL_{50}$ , both methods preferentially recovered the incorrect tree with long terminal branches grouped together. Results were equivalent when a covarion heterotachy model was used ( $BL_{50} = 0.28$ ). MP was also inaccurate under FZH conditions, but it recovered the correct tree with significantly less phylogenetic signal than either model-based technique ( $BL_{50} = 0.22$ ,  $P < 0.001$ ). As previously reported [62], both standard model-based techniques and maximum parsimony are statistically inconsistent under FZH conditions, converging on the incorrect “long branch attraction” topology as sequence length increases at internal branch lengths below the  $BL_{50}$ . The mixed branch length model is significantly more accurate than any existing technique tested ( $BL_{50} < 0.001$ ,  $P < 0.001$ ), confirming previous findings [104]. AIC selected the correct number of branch length sets 96% of the time, and under these conditions, the mixed branch length model was nearly as accurate as when the true evolutionary process was known in advance.



**FIGURE 3.2:** Accuracy of standard model-based phylogenetic inference is impaired by Felsenstein Zone Heterotachy (FZH); the mixed branch length model improves accuracy. **A)** Proportion of replicate datasets from which the correct tree is uniquely recovered by each method is plotted against increasing internal branch length; sequences of 5000 nt were simulated using the tree at left, with long terminal branch lengths of 0.75 substitutions/site and short terminals of 0.05. Models examined were: standard maximum likelihood (ML, blue dots), Bayesian MCMC (BMCMC, red dots), maximum parsimony (gray dots), a Bayesian implementation of the covarion model (red diamonds) and the mixed branch length model (green X's). The true maximum likelihood model that correctly partitions sites *a priori* is indicated by black crosses. Note that standard ML, BMCMC, and the covarion model performed equivalently, so these series overlap. **B)** Branch lengths estimated using standard ML (blue dots), the mixed branch length model (green X's) and the true ML model (black crosses) are plotted against the true simulated internal branch length. Left panel indicates estimated internal branch lengths, while right panel shows estimated terminal lengths; dotted lines indicate perfect correspondence between simulated and estimated branch lengths.



We additionally examined the accuracy of branch length estimates under FZH conditions. As expected, when the true evolutionary process was known in advance,  $ML_{true}$  produced highly accurate estimates of both the internal branch length—which is the same in both classes—and expected terminal branch lengths across all sites (Figure 3.2B). In contrast, standard ML that assumes a single set of branch lengths applies to all sites produced biased estimates of both lengths. In this case, the internal branch length on the true tree is severely underestimated, even when the correct phylogeny is reliably recovered. Expected terminal branch lengths are similarly underestimated, although the degree of underestimation is small and improves as the internal branch length grows. In contrast to standard ML models, the mixed branch length model was not biased, producing highly accurate branch length estimates that were similar to those obtained using  $ML_{true}$ .

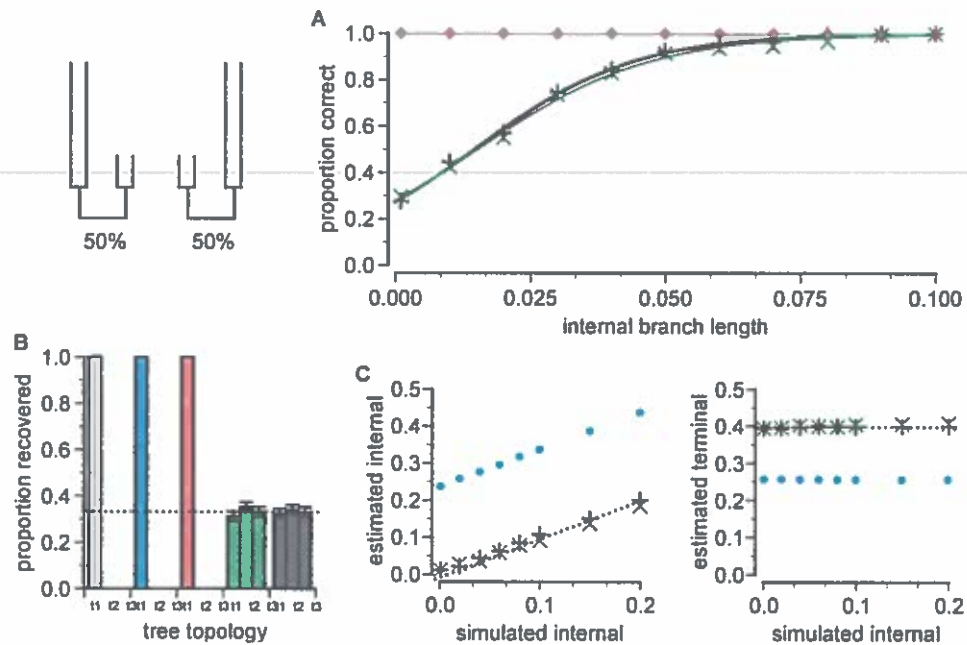
When sequences were simulated under IFZH conditions, with long terminal branches leading to sister taxa, both standard model-based techniques and maximum parsimony recovered the correct phylogeny 100% of the time, even when internal branch lengths were very short and the true evolutionary model failed to recover the correct tree (Figure 3.3A). In contrast, the mixed branch length model (for which AIC correctly inferred two branch length classes 100% of the time) again performed similarly to the true model, exhibiting reduced phylogenetic accuracy as the internal branch length approached zero. In fact, the high accuracy of standard model-based methods and MP in this case was due to a strong directional bias in favor of the long-branch attraction tree, which happened to be correct. To demonstrate this bias, we simulated sequences under IFZH conditions but with a zero-length internal branch. These data were analyzed using standard ML, BMCMC and MP, all of which recovered the incorrect long-branch attraction tree 100% of the time with bootstrap or posterior probability support  $> 0.95$  (Figure 3.3B). In contrast, the true ML model recovered each possible topology with roughly equal proportions ( $P = 0.91$ ), as an unbiased method should [115]. The mixed branch length model performed similarly to the true ML model, recovering each possible tree about the same number of times from replicate datasets ( $P = 0.55$ ). Although we were unable to gauge support for trees inferred using the mixed branch length

model due to the computational demands of bootstrapping a complex model, these results indicate that the mixed branch length model does not suffer from the same topological bias as standard model-based techniques under IFZH.

IFZH conditions also caused standard ML to misestimate branch lengths (Figure 3.3C). While the true model produced accurate internal and terminal branch length estimates, the internal branch length was severely biased upward and expected terminal lengths were underestimated when standard ML was used to analyze IFZH data. The degree of bias was consistent across a range of internal branch lengths. In contrast, the mixed branch length model produced highly accurate estimates of both internal and expected terminal lengths, similar to those using the true model.

### Single Long Branch Heterotachy (SLBH)

Single Long Branch Heterotachy (SLBH) divides sites into four classes; in each class, sites are released from selection in one lineage, but the lineage with an accelerated evolutionary rate differs among classes (see Figure 3.4). As before, we simulated sequences of 5,000 nucleotides under SLBH conditions with long terminal branches of 0.75 substitutions/site and short terminals of 0.05. Equal proportions of sites were assigned to each branch length class. Replicate datasets were generated at a variety of internal branch lengths, and the accuracy of each phylogenetic method was assessed by comparing  $BL_{50}$ s. We found that—similarly to FZH—SLBH caused standard model-based techniques to strongly favor an incorrect phylogeny, resulting in reduced accuracy and statistical inconsistency (Figure 3.4A). While the true evolutionary model produced highly accurate results ( $BL_{50}=0.002$ ), maximum likelihood analysis using the true JC69 model (selected by LRT 92% of the time at  $\alpha=0.05$ ) failed to recover the correct tree at branch lengths below  $BL_{50}=0.015$ . In contrast, maximum parsimony was as accurate as the true ML model under SLBH conditions ( $BL_{50}=0.002$ ). The mixed branch length model ( $BL_{50}=0.003$ ) was significantly more accurate than standard ML, but less accurate than MP or  $ML_{true}$  ( $P < 0.001$ ) when the correct number of branch length classes was assumed. Under

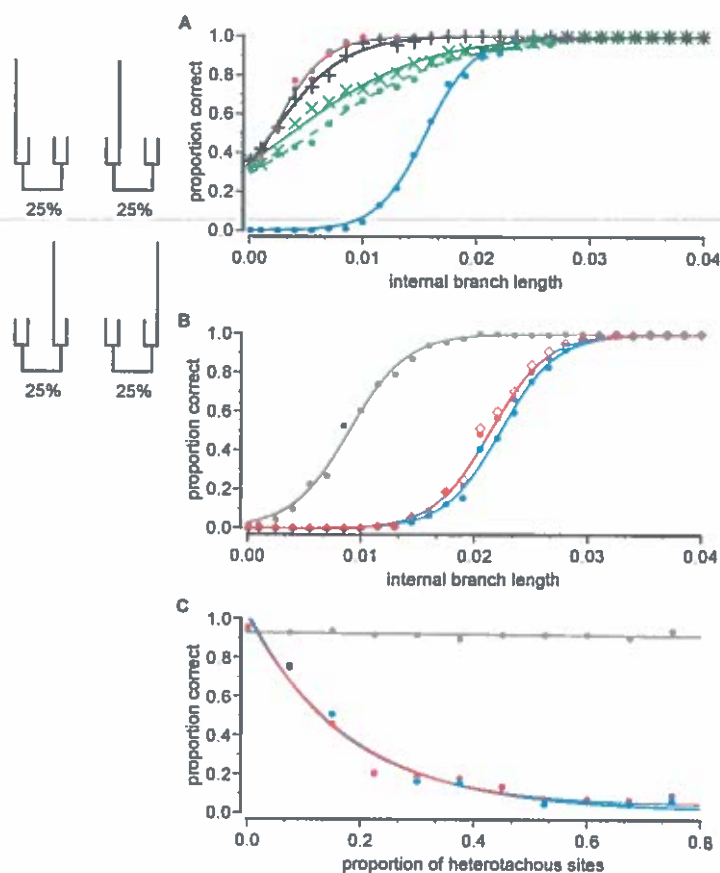


**FIGURE 3.3:** Standard phylogenetic techniques are strongly biased in favor of the correct tree under Inverse Felsenstein Zone Heterotachy (IFZH); the mixed branch length model is unbiased. **A)** Proportion of datasets correctly inferred is plotted against increasing internal branch length when 5000-nt sequences were simulated using the tree at left, with long terminals of 0.75 substitutions/site and 0.05 short terminals. Models examined were: standard maximum likelihood (ML, blue dots), Bayesian MCMC (BMCMC, red dots), maximum parsimony (MP, gray dots), the covarion model (red diamonds) and the mixed branch length model (green X's). The true partitioned model is indicated by black crosses. Note that standard ML, BMCMC, MP, and the covarion model performed equivalently—always recovering the correct tree at all branch lengths examined—so these series overlap. **B)** The proportion of each tree topology (t1,t2,t3) recovered from replicate IFZH datasets simulated with a zero-length internal branch length is shown. Models examined were ML (blue), BMCMC (red), MP (gray), and the mixed branch length model (green), with the true ML model shown in black. Bars indicate standard error; the dotted line indicates equal proportions of each topology recovered. **C)** Branch lengths estimated using standard ML (blue dots), the mixed branch length model (green X's) and the true ML model (black crosses) are plotted against the true simulated internal branch length, with left panel indicating estimated internal and right panel indicating estimated terminal branch lengths. Dotted lines indicate perfect correspondence between simulated and estimated branch lengths.

SLBH conditions, AIC recovered the correct number of data partitions only 50% of the time, and the number of partitions was underestimated as three in 47% of replicates. Assuming three instead of the actual four branch length sets resulted in a slight reduction in accuracy ( $BL_{50}=0.006$ ), but the mixed branch length model was still more accurate than standard ML. While BMCMC appeared highly accurate ( $BL_{50}=0.002$ ), when we required  $> 95\%$  bootstrap or posterior probability support to consider a phylogeny correctly resolved (Figure 3.4B), ML and BMCMC performed nearly equivalently ( $BL_{50}=0.021$  for BMCMC,  $0.022$  for ML), while MP was significantly more accurate than either method ( $BL_{50}=0.009$ ,  $P < 0.001$ ). Use of the covarion model did not improve the accuracy of BMCMC when strong support was required ( $BL_{50}=0.021$ ). Due to the computational demands of the mixed branch length model, we were unable to test the accuracy of this model using bootstrap support.

Even when a very small proportion of sites are heterotachous, the accuracy of standard model-based techniques can be severely reduced by SLBH. We simulated sequences under SLBH conditions with an internal branch length of 0.015, but with a varying proportion of sites in a single branch length class; the remaining sites were equally distributed among the other three classes. Under these conditions, MP recovered the correct tree with  $> 95\%$  bootstrap support about 92% of the time, regardless of the amount of heterotachy present (Figure 3.4C). In contrast, the accuracy of both ML and BMCMC declined with increasing heterotachy. When no heterotachy was present (all sites had the same branch lengths), both methods recovered the correct tree with  $> 95\%$  support 96% of the time. However, when only 7.5% of the sites had different branch lengths, ML recovered the correct tree with strong support only 76% of the time, and BMCMC recovered the correct tree 75% of the time with posterior probability  $> 0.95$ . When half the sites were heterotachous, ML and BMCMC recovered the correct tree with strong support from fewer than 15% of replicates.

The reason standard model-based techniques perform poorly under SLBH conditions is that both the internal branch length and expected terminal branch lengths are severely underestimated (Figure 3.5A); at internal branch lengths below

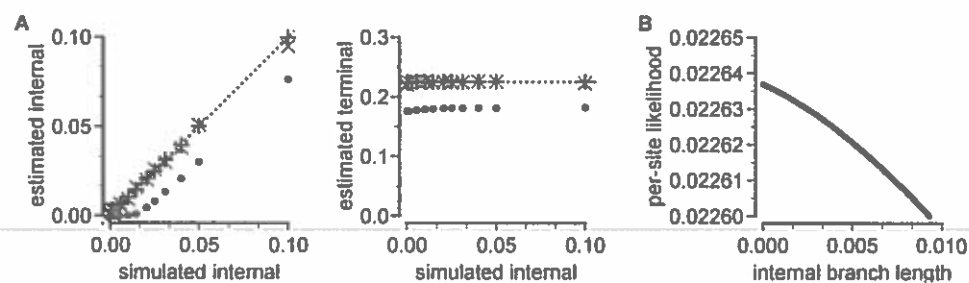


**FIGURE 3.4:** Single Long Branch Heterotachy (SLBH) impairs the accuracy of standard model-based phylogenetic techniques; the mixed branch length model improves accuracy. **A)** Proportion of correct inferences is plotted against increasing internal branch length for various models when 5000-nt sequences were simulated using the tree at left; long terminal branch lengths were 0.75 substitutions/site, while short terminals were 0.05. Models were: standard maximum likelihood (ML, blue dots), Bayesian MCMC (BMCMC, red dots), maximum parsimony (MP, gray dots), and the mixed branch length model with either the correct number of branch length categories (green X's) or only three categories (green dots). True partitioned model is indicated by black crosses. **B)** Proportion of correct inferences with support > 0.95 is plotted against increasing internal branch length, with support measured by non-parametric bootstrapping for ML and MP and by posterior probability for BMCMC. Red diamonds indicate Bayesian covarion model. **C)** Proportion of correct inferences with > 0.95 support is plotted against increasing proportion of heterotachous sites. Sequences were simulated under SLBH conditions with an internal branch length of 0.015, but the proportion of sites in branch length classes 2-4 (x-axis) varied from zero (no heterotachy) to 0.75.

the BL<sub>50</sub>, the maximum likelihood estimate of the internal branch length on the best tree is zero, resulting in an inference of an unresolved tree. Even when the internal branch length is long enough for ML to reliably recover the correct tree, the branch length is substantially underestimated—as are expected terminal lengths. In contrast to standard ML, the mixed branch length model produces highly accurate estimates of both the internal and expected terminal branch lengths, resulting in improved phylogenetic accuracy. The inference of a zero-length internal branch by standard ML does not improve with increasing sequence length, resulting in statistical inconsistency at branch lengths below the BL<sub>50</sub>. Even with infinite sequence data, ML inferred a zero-length internal branch from SLBH data with a true internal branch length of 0.01 (Figure 3.5B); we calculated the per-site likelihood of an infinite SLBH dataset at various internal branch lengths, and found that a branch length of zero produced the highest likelihood. These results indicate that standard ML is statistically inconsistent under SLBH conditions. Integrating over multiple internal branch length values using BMCMC results in recovery of the correct phylogeny, because phylogenetic signal favors the true tree when a non-zero internal branch length is imposed. However, support for the true tree is very weak using standard BMCMC, because SLBH conditions produce the greatest likelihood when the internal branch length is zero.

### Signal-Noise Heterotachy (SNH)

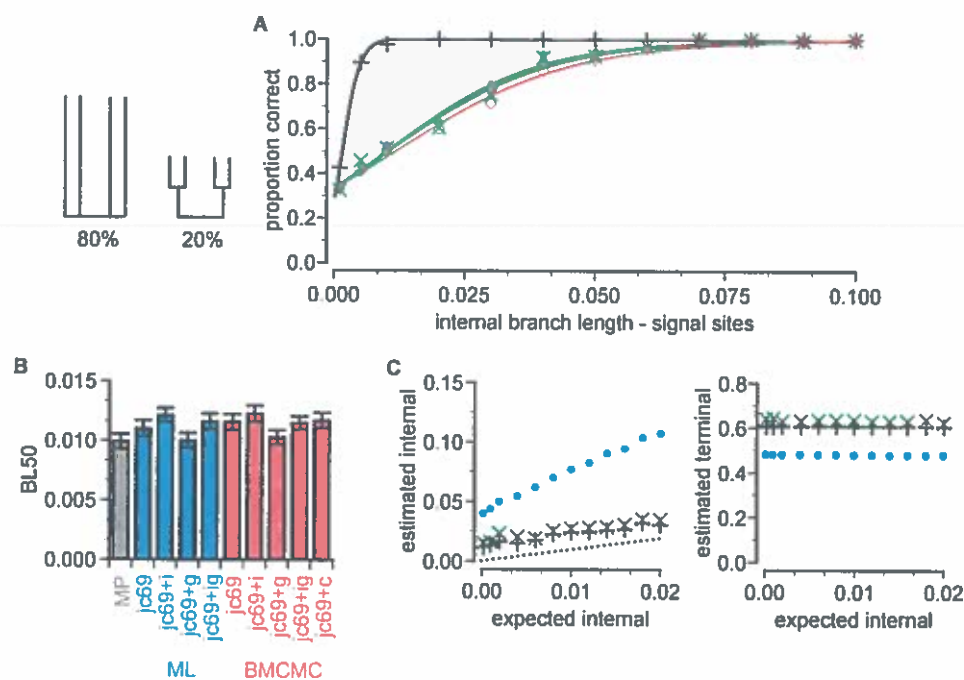
In Signal-Noise Heterotachy (SNH), sites are divided into two categories. In one category, “noisy” sites are essentially randomized, with long terminal branch lengths (0.75 substitutions/site) and no internal branch; in the other category, sites evolve with short terminal branch lengths (0.05) and a variable internal branch (see Figure 3.6). We simulated 5,000-character SNH datasets with various proportions of noisy sites, analyzed these data using the methods described above, and compared the accuracy of different methods by comparing BL<sub>50</sub> estimates. Consistent with previous findings that random sequence segments impair phylogenetic accuracy [111], we found that SNH caused a reduction in the accuracy of phylogenetic



**FIGURE 3.5:** Standard maximum likelihood misestimates branch lengths under Single Long Branch Heterotachy (SLBH, see Figure 3.4); the mixed branch length model provides improved estimates. **A)** Branch lengths estimated by standard maximum likelihood (ML, blue dots), the mixed branch length model (green X's), and the true partitioned ML model (black crosses) are plotted against the true simulated internal branch length. Left panel shows estimated internal branch length, while right panel indicates terminal branch lengths; dotted line indicates perfect correspondence between estimated and actual branch lengths. **B)** Per-site likelihood calculated on an ideal infinite dataset (see Materials and Methods) using standard ML is plotted against increasing internal branch length; sequence was generated under SLBH conditions with long terminals of 0.75 substitutions/site, short terminals of 0.05, and a true internal branch length of 0.01.

inference for all methods tested (Figure 3.6A). While the true model ( $ML_{true}$ ) produced highly accurate phylogenetic estimates at very short internal branch lengths, phylogenetic accuracy was reduced when the true evolutionary process was not known in advance ( $P < 0.001$ ). The likelihood ratio test for the best-fit model was ambiguous in this case, selecting the JC69+G+I model 38% of the time, the JC69+I model 32% of the time, and the JC69+G model 28% of the time. The JC69+G model was the most accurate, but differences in accuracy among the various models (including the covarion model) were small (Figure 3.6B). The mixed branch length model did not substantially improve the accuracy of phylogenetic inference under SNH conditions, although AIC selected the correct number of branch length categories 99% of the time.

Branch length estimates were biased using standard ML, which severely overestimated the expected internal branch length while underestimating terminal lengths (Figure 3.6C). In contrast, both  $ML_{true}$  and the mixed model accurately



**FIGURE 3.6:** Signal-Noise Heterotachy (SNH) impairs the accuracy of phylogenetic inference. **A)** The proportion of trees correctly inferred is plotted against increasing internal branch length for sites containing phylogenetic signal (20% of sites, with terminal branch lengths 0.05 substitutions/site); the remaining sites (80%) are randomized (terminal branch lengths 0.75) with no phylogenetic signal. Sequences were 5000 nt long. Results are shown for 80% noise; similar results were obtained when the proportion of noisy sites varied from 70% to 95% (not shown). Methods examined were: standard maximum likelihood (ML, blue dots, jc69+ig model), Bayesian MCMC (BMCMC, red dots), maximum parsimony (MP, gray dots), a Bayesian covarion model (red diamonds), and the mixed branch length model (green X's). Black crosses indicate performance of the true partitioned ML model. Note that ML, BMCMC, MP, and the mixed branch length model all performed similarly, so these series overlap. **B)** The internal branch length at which 50% correct recovery was obtained is shown for a variety of evolutionary models implemented in either ML (blue) or BMCMC (red). The invariant sites model is indicated by +i, the gamma model by +g, and the covarion model by +c. Note that a lower BL<sub>50</sub> indicates a more accurate method. **C)** Internal branch lengths estimated using standard ML (blue dots), the mixed branch length model (green X's) and the true ML model (black crosses) are plotted against the true expected internal branch length (averaged over both noisy and signal sites). Left panel shows estimated internal branch length, while right panel shows estimated terminal branch lengths; dotted line indicates perfect correspondence between estimated and actual expected branch lengths.



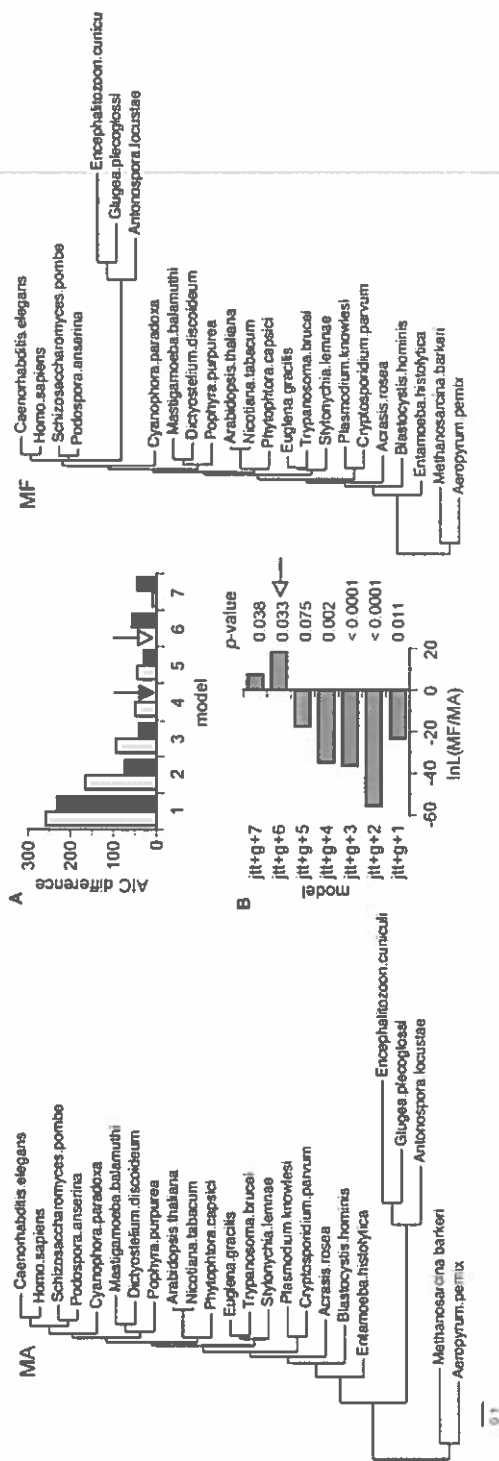
estimated expected terminal lengths, although expected internal branch lengths were slightly overestimated by both methods. The reason the internal branch length on the true tree was overestimated by the true model is that stochastic error inflates the estimate of the internal branch length for noisy sites, which is actually zero: when stochastic error in noisy sites produces state patterns that happen to favor the correct tree, the internal branch length for noisy sites is overestimated. When stochastic error favors an incorrect tree, the internal branch length for noisy sites on the correct tree is inferred as zero. As a result, the net effect is to overestimate the internal branch length for noisy sites—and hence the expected internal branch length.

### Summary

Taken as a whole, the results of our heterogenous branch length simulations show that various forms of heterotachy can negatively affect the accuracy of phylogenetic inferences when standard evolutionary models are employed. The mixed branch length model provides much more accurate inferences of both phylogeny and evolutionary model parameters than standard techniques under simulated heterotachy. Although the mixed branch length model can sometimes perform nearly as well as the true partitioned model (FZH and IFZH), in other cases there is a significant loss of accuracy when the mixed model is compared to the true model (SLBH and SNH).

### 3.3.2 Elongation Factor $1\alpha$ Sequences

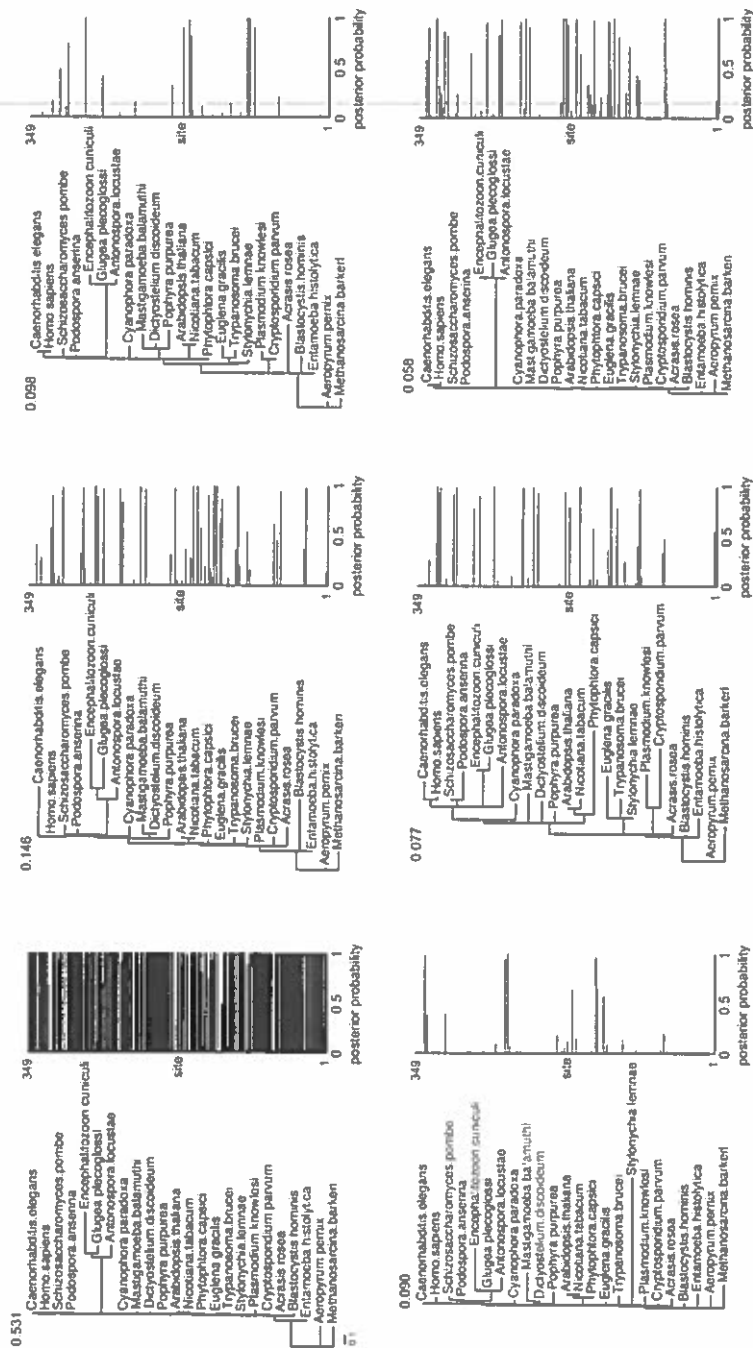
Although simulated evolution can establish the potential performance impacts of heterotachy on phylogenetic inference, the true test of any method is how accurately can it reconstruct correct evolutionary relationships from actual sequence data. To determine whether the mixed branch length model can infer accurate phylogenies from data that confound conventional methods, we analyzed the elongation factor  $1\alpha$  dataset of Inagaki et al. [55], who showed that heterotachous “rate shifts” cause both MP and ML to artifactually group Microsporidia with the Archaeobacterial



**FIGURE 3.7:** A mixed branch length model recovers the correct Microsporidia + Fungi (MF) grouping from elongation factor 1 $\alpha$  sequence data. Correct MF tree is shown at right, and incorrect Microsporidia + Archaeobacteria (MA) tree is shown at left, with branch lengths inferred by maximum likelihood using the jtt+g model. **A)** The difference in AIC scores between each model and the model with minimal AIC is plotted for the jtt+g model with 1-7 branch length classes. Dark bars indicate AIC scores calculated using the MA tree, while light bars indicate those calculated using the MF tree, with corresponding arrows indicating the model selected by AIC assuming each topology. **B)** The log-likelihood ratio of the MF tree to the MA tree is plotted for models with increasing number of branch length classes, with negative lnL ratios indicating support for the MA tree and positive values indicating support for the MF tree. The significance of support for the best tree in each case was assessed using the AU test. Arrow indicates the model selected by AIC assuming the correct MF topology.

outgroup (the MA tree) rather than correctly with Fungi (MF, see Figure 3.7). First, we analyzed Inagaki et al's Micro\* dataset using unweighted MP and ML using the JTT+gamma model, confirming that both methods recover the artifactual MA tree (not shown). Bayesian analysis using the covarion model also recovered the MA tree with weak support (posterior probability 0.58, not shown). In contrast, the mixed branch length model recovered the correct MF topology with significant support when the number of branch length categories was selected using AIC (Figure 3.7). We calculated the AIC value for each number of branch length categories, from one (the simple JTT+gamma model) to seven, using both MF and MA topologies; while the best-fit number of categories calculated using the correct MF tree was six, using the incorrect MA phylogeny caused AIC to underestimate the amount of heterotachy, resulting in only four branch length categories (Figure 3.7A).

Underestimating the correct number of branch length categories resulted in erroneous support for the MA tree (Figure 3.7B), while the correct MF tree was supported when six or seven branch length categories were used. We used the AU test [99] to determine whether the maximum likelihood tree was significantly supported and found that when the AIC-selected number of branch length categories (six) was used, the correct MF tree was weakly but significantly supported over the artifactual MA grouping ( $P = 0.033$ ). To assess whether the mixed branch length model can partition sites into branch length classes, we calculated the posterior probability of each branch length class, given each site in the dataset (see Materials and Methods, Figure 3.8). We found that sites producing high posterior probability existed for each of the six branch length classes, with the number of sites giving high posterior probability generally following the weights inferred for each branch length category from the data. The ability of the mixed branch length model to partition sites among its inferred branch length categories suggests that the model is indeed capturing an important part of the underlying evolutionary process. Taken as a whole, these elongation factor  $1\alpha$  results show that the mixed branch length model can recover correct evolutionary relationships from difficult data, providing improved phylogenetic accuracy compared to existing methods under challenging real-world conditions.



**FIGURE 3.8:** Mixed model analysis of elongation factor 1 $\alpha$  data partitions sites into branch length categories. We plot the posterior probability that each site in the alignment evolved according to each set of branch lengths inferred using a 6-category mixed branch length model (inferred branch lengths shown to the left of each graph). The number above each tree indicates the inferred proportion of sites expected to evolve according to those branch lengths.

### 3.4 Discussion

Accurately inferring evolutionary relationships and parameters describing the evolutionary process from molecular sequence data is a challenging but vital problem in evolutionary biology. Evolutionary conditions that strongly violate the assumptions of phylogenetic models can result in biased and inaccurate inferences, providing a misleading picture of how life has evolved. Site-specific changes in evolutionary rates (heterotachy) have been widely documented in empirical sequence data [5, 31, 48, 69, 71, 72, 75, 76, 77, 81, 84, 91], but models of heterotachy are not regularly used for phylogenetic analyses. We have shown here that various forms of heterotachy can cause standard evolutionary models to infer inaccurate phylogenies, sometimes with strong support. Furthermore, accuracy is not always improved by increasing the amount of sequence data available; under some heterotachous conditions, model-based techniques would infer incorrect trees even if infinite data were available (see also [11, 62]). In some cases, the nonparametric technique maximum parsimony is significantly more accurate than maximum likelihood and Bayesian methods using commonly-available evolutionary models. Although the covarion model did not produce more accurate phylogenies than homotachous models under the conditions we examined, a mixed branch length model allowing different sites to evolve along different branch lengths did improve the accuracy of phylogenetic inferences, both from simulated and real-world data.

We have also shown that AIC can be an effective technique for determining the best-fit number of branch length categories for a mixed branch length model analysis. In contrast, our additional analyses (not shown) show that both the corrected AIC (AICc) and BIC underestimate the number of branch length categories. For example, BIC always supported a simple model with a single set of branch lengths from four-taxon datasets simulated with multiple branch lengths (FZH, IFZH, SLBH, SNH), even when ample data were available (5,000 nt) and the simple model was significantly biased. Similarly, AICc underestimated the number of branch length categories from elongation factor  $1\alpha$  data, preferring a model with only two categories of branch lengths which strongly supported the incorrect MA

tree; AIC selected a six-category model that resulted in significant support for the correct tree. In contrast to AIC, which applies a constant penalty for additional model parameters, both AICc and BIC penalize the addition of model parameters proportionally to the number of sites in the sequence [88]; our analyses suggest that this is not the best approach for selecting the number of branch length categories for a mixed model. Likelihood Ratio Tests (LRTs) cannot be easily used to select the best-fit number of branch length categories, as the LRT statistic is not chi-square distributed in this case [73].

Although the mixed branch length model improved the quality of phylogenetic inferences in our study, the computational complexity of the model remains a nagging concern. Since likelihoods must be computed separately for each set of branch lengths in the model, the number of required likelihood calculations increases quickly with increasing evolutionary heterogeneity. In the case of elongation factor  $1\alpha$  sequences, likelihood calculations using gamma-distributed among-site rate variation with four rate categories and heterotachy with six branch length categories required 24 separate likelihood calculations, compared to only four using the simpler gamma model with a single set of branch lengths. In addition to the increase in the number of likelihood calculations, the increased number of parameters in the mixed model requires additional computation time to estimate. The complex interactions and constraints on mixed branch length parameters make simple hillclimbing heuristics ineffective, requiring slower but more robust techniques such as simulated annealing to reliably estimate evolutionary parameters under the complex model. The dramatic increase in computational resources required to analyze data using a mixed branch length model prevents calculation of common support measures such as bootstrapping and could limit the number of taxa or amount of sequence data that can be analyzed in a reasonable amount of time. Developing techniques to accurately and quickly solve mixed branch length models and estimate statistical confidence in phylogenies inferred using mixed branch length techniques is clearly an important area for future research. Because the mixed model can produce accurate phylogenies under conditions that mislead existing methods, however, mixed branch length analysis can be used to assess the

robustness of existing inferences to model violations such as heterotachy. The ability to infer site-specific evolutionary properties by calculating the posterior probability that each site evolved along each set of branch lengths offers a novel and potentially insightful window into the complexity of the molecular evolutionary process.

## CHAPTER IV

# OPTIMIZATION OF MIXED BRANCH LENGTH MODELS

## 4.1 Computational Challenges in Phylogenetic Inference

Typical phylogenies produced today are inferred from molecular sequence data using complex evolutionary models and sophisticated statistical estimation techniques. It is not uncommon for published trees to have on the order of 50 taxa [58, 79], and sequence data sets with tens of thousands of characters are becoming increasingly used to resolve difficult problems [83, 94]. When data are combined from multiple genes extracted from a wide variety of organisms spanning many large taxonomic groups, the process governing molecular evolution on this scale is likely to be very complex, requiring a complicated evolutionary model in order to capture important features of the process and avoid the potential for errors caused by model underparameterization [9, 11, 55, 62, 131]. Inferring a large phylogenetic tree from long molecular sequences using a complex evolutionary model is a computational challenge; careful implementation of efficient algorithms is required to balance the requirements of high accuracy and high speed.

It has been recognized for some time that increasing the number of taxa under study can have a beneficial effect on phylogenetic accuracy [43]. As the number of taxa increases, however, the number of possible trees grows exponentially, requiring



heuristic tree-search algorithms to identify the optimal topology. Current implementations rely almost exclusively on simple hillclimbing heuristics: given a topology at iteration  $i$ , a new topology ( $i + 1$ ) is proposed by rearranging some of the nodes of tree  $i$ . This proposal is accepted if it improves the optimality score of the tree, otherwise it is rejected. The procedure is repeated until some stopping criterion is reached, typically either no improvement in score for  $k$  consecutive iterations or a pre-set iteration limit. Although these techniques work quite well for moderately-sized topologies, they suffer from two major drawbacks that could limit their effectiveness for inferring very large trees. First, most heuristic search strategies in a maximum likelihood (ML) framework require the optimization of all model parameters at each iteration step [98]. If the model is complex, parameter optimization can be costly, thereby limiting the efficiency of the search algorithm. Some algorithms have attempted to circumvent this issue by reducing the complexity of tree rearrangements and the stringency of parameter optimization [37, 106, 121], which increases the efficiency of the algorithm but may reduce the rigor with which tree space is searched and parameter optimization is done, potentially leading to phylogenetic errors [9]. The second potential problem with hillclimbing is the well-known issue of the algorithm becoming stuck in suboptimal regions of parameter space. Although very little is known about the functional landscape of phylogenetic problems, recent analyses suggest that real-world phylogenetic problems may contain multiple peaks and valleys, potentially causing hillclimbing techniques to become stuck in local optima [9, 61]. These problems could become much worse as the complexity of the problem increases.

Recent advances in Bayesian phylogenetics have attempted to entirely sidestep the computational issues raised when trying to find the best tree by focusing instead on estimating a 'credible set' of trees with high likelihood using Markov Chain Monte Carlo (MCMC [51]). MCMC uses the same iterative proposal mechanism as hillclimbing ML but extends this technique to estimating model parameters as well as topology. Rather than attempting to discover the best tree and parameter values, Bayesian techniques use MCMC to 'integrate over' all possible trees and parameters. The result is that relatively expensive numerical estimation techniques

for parameter optimization are not necessary; the simple proposal mechanism can be extended to sample model parameters as well as topologies. At each step in the algorithm, a new topology and/or set of model parameters is proposed by slightly altering the existing values; these proposals are either accepted or rejected, with the acceptance probability being proportional to the likelihood ratio of the proposed state to the old state. The algorithm continues until a pre-set number of proposals have been attempted. At given intervals, the state of the MCMC run is recorded, and the relative frequency with which each tree is sampled gives an estimate of the tree's posterior probability. Integrating over trees and model parameters using MCMC results in a much faster algorithm than parameter optimization by ML hillclimbing but leads to a new set of potential problems. First, although MCMC is theoretically sound, concerns have been raised that it may not always be accurate in practice, especially if the phylogenetic problem is very complex [78, 93]. Second, recent concerns have also been raised that even under simple simulation conditions when the correct evolutionary model is used, Bayesian methods may overstate the statistical confidence in the best tree, sometimes leading to high rates of error [12, 17, 67, 74, 102, 112, 117].

The computational issues associated with molecular phylogenetics increase dramatically when the evolutionary model is made more complex. We recently introduced a mixed branch length model to account for heterogeneous evolution [62, 104] and showed that this model can improve the accuracy of inferred phylogenies using both simulated and empirical data (chapter 3). The main drawback to this model is the potential explosion in the number of parameters; not only are there more parameters, but complex interactions among different parameters make fast numerical optimization impossible. Instead, we implemented a simulated annealing algorithm to optimize model parameter values. Although this algorithm provided highly accurate parameter estimates under simulation conditions, the time required to perform the calculations was prohibitive. As a result, the general usefulness of the mixed model for phylogenetic inference is questionable.

Here we address the computational challenges associated with our implementation of the mixed branch length model. We develop strategies to improve the efficiency of our software by addressing the problem from multiple levels, including parallelization, algorithmic improvement, and code optimization. First, we use runtime performance analysis to identify code ‘bottlenecks’ and propose restructuring the code to address these issues. Second, we identify faster simulated annealing algorithms that can be used to speed up our implementation without sacrificing optimization rigor. Third, we outline an approach using simulated annealing to optimize tree topology, parameter values, and model complexity simultaneously, reducing the need to perform multiple independent runs using different trees and/or models. Finally, we identify potential avenues for parallelization and discuss the pros and cons of each. Using a combination of multiple strategies to improve the efficiency of our phylogenetic reconstruction algorithm should result in the necessary runtime improvements required to make mixed-model phylogenetic analysis feasible for general use. Many of the improvements outlined here should improve the running time of simpler models as well, providing a very general strategy for fast, accurate phylogenetic reconstruction.

## 4.2 Why Simulated Annealing?

A simple simulated annealing algorithm was used to optimize model parameters for the mixed branch length model implemented in chapter 3. Although slower than numerical techniques such as simplex and quasi-Newton methods, simulated annealing is widely used to optimize complex functions. The major advantages of simulated annealing over faster numerical methods are 1) ease of programming, 2) applicability to constrained optimization problems, and 3) an ability to avoid local optima. While numerical methods typically require calculation of derivatives in order to quickly converge to a local optimum, simulated annealing uses a simple stochastic proposal mechanism that requires no derivative information. Due to this simple proposal mechanism, simulated annealing can be easily applied to highly constrained optimization problems, while many numerical methods are only

applicable to unconstrained parameters. The mixed branch length model has many complex constraints on model parameters. For example, branch lengths must be at least 0; the proportion of sites in each branch length class must also be  $\geq 0$ , and the sum of site proportions must be 1.0. Finally, the acceptance of ‘bad’ proposals that reduce the likelihood score allows simulated annealing to effectively traverse areas of low likelihood to find function peaks, thereby avoiding local optima. Numerical optimization algorithms rely on hillclimbing and can easily become stuck if multiple local optima exist. Because evidence suggests that both phylogenetic problems and mixed models may generate complex functional landscapes with multiple local optima [61, 73], hillclimbing methods may not be appropriate for optimizing the mixed branch length model.

### 4.3 Code Optimization

The prototype mixed branch length model—optimized using simulated annealing—was sufficient to address moderate-sized phylogenetic problems such as the 24-taxa, 349-character elongation factor  $1\alpha$  data set of Inagaki et al. [55] analyzed in chapter 3. Analysis of these data using the best-fit 6-category model required about 2 weeks of computing time on a 2-gigahertz G5 processor running mac OS X. Already pushing the patience of most biologists, analyses of larger data sets was even more daunting. A recent analysis of a 92-taxon, 1448-character dinoflagellate data set required over 2 months to complete, and the 49-taxa, 35,371-character bilaterian data set of Philippe et al. [83] crashed before completing. A 183-taxa, 438-character nuclear receptor data set also failed to complete. These results suggest that the existing prototype software may require significant improvements before it can be applied to large phylogenetic problems.

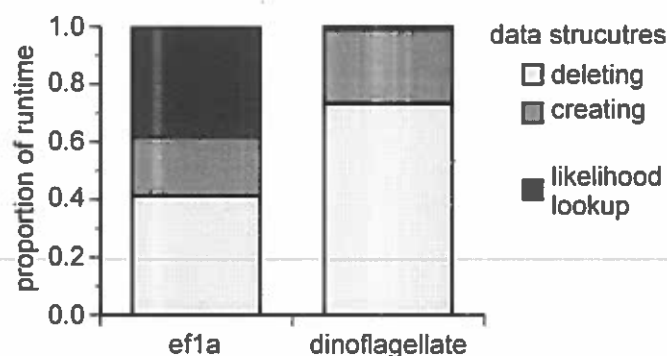
In order to target code optimizations to parts of the software likely to produce the greatest gains in efficiency, we used runtime profiling to identify the subroutines responsible for most of the program’s execution time. Runtime profiling was performed using Shark v4.3, included as part of the CHUD tools with mac OS X. Because data sets with different properties might tax different parts of the software,

we profiled the software while analyzing three separate data sets: 1) the elongation factor  $1\alpha$  ( $ef1\alpha$ ) data analyzed in chapter 3 [55], 2) the dinoflagellate data set described above, and 3) the concatenated bilaterian data set [83]. The  $ef1\alpha$  data was moderate in terms of both taxon sampling and sequence length; the dinoflagellate data sampled many more taxa but still relatively few characters from each taxon, while the bilaterian data had a moderate number of taxa but very long sequences.

### 4.3.1 Elongation Factor $1\alpha$ ( $ef1\alpha$ ) Data Profiling

Runtime analysis of the mixed branch length model software applied to the elongation factor  $1\alpha$  ( $ef1\alpha$ ) data revealed that 99.7% of function samples were within the simulated annealing routine, so overhead required to read and process input files or write output files was negligible. Within the simulated annealing routine, 61.6% of the total time was spent copying the model from one memory location to another, while 38.1% of runtime was spent calculating likelihood scores (Figure 4.1). When the simulated annealing algorithm makes a new proposal, the entire data structure encoding the model—including the tree with multiple branch lengths, transition model parameters, etc.—is copied into a new memory location. The copy is then changed slightly, its likelihood calculated, and the likelihood of the new proposal is compared to that of the old state. If the proposal is accepted, the old data structure is deleted and replaced by the new data structure. Deletion of data structures accounted for 41.3% of total runtime, while allocation of new data structures required 20.3% of runtime.

Within the likelihood calculation, nearly all of the time was spent looking up transition probabilities (38.0% of total runtime); only 0.1% of total runtime was required to actually calculate transition probability matrices. Once the transition probability matrices have been built, the likelihood calculation consists of a complex series of nested loops that combine various transition probabilities to calculate the total likelihood using a post-order tree traversal [22]. Interestingly, these loops required much more processor time than the matrix operations required to calculate transition probabilities.



**FIGURE 4.1:** Runtime profiling reveals data structure copying as the major performance bottleneck. The proportion of total runtime spent in various procedures is plotted when elongation factor 1 alpha (ef1a) and dinoflagellate data sets were analyzed using our prototype mixed branch length model software. Data structure manipulations are shown in gray, while time spent looking up likelihood values is shown in black.

### 4.3.2 Dinoflagellate Data Profiling

The dinoflagellate data set has many more taxa than the ef1 $\alpha$  data set (98 vs. 24), but only moderately increased sequence length (1448 vs. 349 characters). When the software was profiled while analyzing the dinoflagellate data, all of the samples were within the simulated annealing routine, again indicating that file input/output overhead was negligible (Figure 4.1). As with the ef1 $\alpha$  data, nearly all of the processor time was spent either copying data structures from one place to another or calculating likelihood scores. However, the partitioning of time between these two operations was very different in the case of the 98-taxa dinoflagellate data. In this case, 98.1% of total runtime was spent copying data structures, while only 1.1% of total time was spent actually calculating likelihood scores. Of the time spent copying data, 73.4% of total runtime was required to delete various data structures, while 25.5% was required to create new structures. All of the samples taken from within the likelihood calculation routines were transition probability lookups rather than matrix operations.

### 4.3.3 Bilaterian Data Profiling

The bilaterian data consisted of relatively few taxa (49) but very long sequences (35,371 characters). Under these conditions, the prototype software failed to complete reading the sequence data into memory. Most of the total runtime (86.1%) was devoted to system calls, while only 12.8% of the function samples were actually taken within the main program execution. Analysis of the runtime trace confirmed that the sequence data were never completely read into memory. We used the nexus class library (NCL v2.0 [66]) to read input data files, and this library was insufficient to handle the very large bilaterian data set. All of the functions sampled within the main program execution were NCL library functions.

### 4.3.4 Planned Optimizations

To address the optimization issues identified by runtime profiling, we plan to refactor the prototype software as follows. First, it is obvious from the bilaterian data analysis that the nexus class library used to read input data files is not efficient enough to analyze extremely long sequences. Because such 'genomic' data sets are becoming increasingly exploited to address difficult phylogenetic problems, we will replace the nexus class library with more efficient data input routines. Second, we plan to explore the potential of more aggressive loop optimizations to reduce the time required to calculate likelihood scores. Finally, it is obvious from the *efl* $\alpha$  and dinoflagellate data sets (Figure 4.1) that the copying of data structures as part of the simulated annealing loop must be streamlined in order to improve the performance of the algorithm. Our initially naive approach of creating a new data structure and deleting the old data structure at each iteration is clearly inadequate.

To address this problem, we plan to implement a 'reversible' proposal mechanism to entirely avoid data structure copying and deleting. Rather than copying the entire data structure into a new memory location, we will instead record the specific parameters changed during each iteration as well as their original values. Parameters will then be altered on the original data structure and the likelihood of the new proposal will be compared to that of the old state as before. If

the proposal is rejected, the altered parameters will be reset to their original states. This scheme should completely eliminate the need to create and delete large data structures as part of the simulated annealing algorithm; runtime profiling analysis suggests that the potential performance gains are considerable.

## 4.4 Faster Simulated Annealing

Although simulated annealing is widely used to optimize complex functions, it is typically slower than numerical estimation techniques, and so is applied in cases where numerical estimation is not appropriate. One of the main advantages of simulated annealing over numerical optimization—the ability to avoid local optima—is also one of the reasons simulated annealing is slow. Simulated annealing relies on a ‘stochastic walk’ through parameter space to sample the function being optimized. As this ‘walk’ is being conducted, the *temperature* parameter is slowly decreasing, reducing the probability of accepting steps that reduce the function value—and thus the ability of the algorithm to walk out of local optima. If the temperature is decreased too rapidly, the algorithm can easily become stuck in a locally optimal region of parameter space. This requirement for a slow temperature reduction regime (called the “annealing schedule”) is the main cause of simulated annealing’s relatively long running time.

Because the annealing schedule is crucial for both the efficiency and accuracy of simulated annealing, considerable attention has been invested into determining ideal methods for temperature reduction. Although many aspects of the annealing schedule are problem specific, some general approaches have been developed. Although it has been proven that a logarithmic annealing schedule—where the temperature at step  $i$  is given by:  $T_i = \frac{R}{\log(i)}$ —will converge to the global optimum given large enough  $R$  [33], this schedule is typically far too slow to be useful in practice. We implemented the most common annealing schedule, an exponential descent where  $T_i = T_{i-1} * a$ . Here  $a$  is smaller than but very close to 1.0. Unfortunately, it is well known that this schedule is not guaranteed to converge on the globally optimal solution [38], and it is still quite slow. Ingber has shown,



however, that an exponential schedule can safely be used provided specific proposal mechanisms are employed [56]. Faster annealing strategies rely almost exclusively on adaptive approaches that make use of statistical sampling of the algorithm to ‘tune’ the annealing schedule as the algorithm proceeds [56, 64]. The schedule proposed by Lam and Delosme [64] estimates the local standard deviation of the function being optimized and either accelerates or decelerates the annealing schedule based on whether the standard deviation is small or large, respectively. The approach used by Ingber [56] estimates the partial derivatives of the function to optimize and uses this information to alter the current temperature independently for each parameter; the temperature is increased if the partial derivative of a parameter is relatively small, otherwise it is decreased.

To improve the quality of our simulated annealing software, we will implement Ingber’s proposal mechanisms to guarantee that the exponential annealing schedule will converge to the global optimum. We will additionally explore the potential for adaptive annealing strategies to further improve the running time of our implementation by making use of algorithmic sampling to intelligently alter the program’s execution as it proceeds.

## 4.5 Model Selection Using Simulated Annealing and Akaike Information Criterion (AIC)

In the case of empirical data, the appropriate number of branch length sets to use in a mixed branch length model analysis will almost never be known in advance. Therefore, estimating the best-fit number of branch length sets is a crucial aspect of the analysis. In chapter 3, we used the Akaike Information Criterion (AIC, [1]) to select the best-fit number of branch length sets, and this criterion was typically very accurate under simulation conditions. A downside to this approach is the computational cost incurred: in our prototype software, several analyses were conducted separately on the same data, each using a different number of branch

length sets. Only after all the analyses had completed was the best-fit model selected using AIC.

An alternative to this computationally intensive approach is to optimize model complexity directly as part of the simulated annealing algorithm. In addition to the typical model parameter and topology proposals, a new type of proposal would either increase the complexity of the model—by adding an additional set of branch lengths to the tree—or decrease the number of branch length sets by merging two sets into one. These ‘reversible jumps’ in model complexity would allow for simultaneous optimization of topology, model parameters, and the complexity of the model. This approach should free up computer resources by focusing the algorithm’s execution on models with good statistical fit to the data rather than having to serially optimize several models, many of which may be either grossly simplistic or overly complex.

A similar ‘reversible jump’ strategy has been successfully used in Bayesian MCMC algorithms to select model complexity [2, 4], including one implementation used to select the complexity of the relative transition rate matrix for phylogenetic analysis [50]. In a Bayesian setting, model complexity need not be penalized statistically, because integration over additional parameters in more complex models automatically penalizes added complexity. In a maximum likelihood approach, however, increasingly complex models will almost always produce higher maximized likelihoods, so models with additional parameters must be penalized to find the best-fit model without overfitting stochastic variation in the data. The AIC and other model-selection criteria include a ‘penalty term’ that accounts for additional parameters in more complex models, so models of various complexity can be directly compared. If the simulated annealing algorithm were used to optimize AIC instead of directly optimizing the likelihood score, the same reversible jump strategy used in Bayesian MCMC could be utilized in a maximum likelihood framework to determine the best-fit model and optimize model parameters and tree topology as part of a single simulated annealing run.

## 4.6 Parallel Algorithms

The advent of fast ‘supercomputers’ created by connecting standard machines into a cluster of interacting nodes has brought high-performance computing to practicing scientists, and the most recently developed phylogenetic inference packages make use of parallel algorithms to exploit cluster computing resources. Unfortunately, the simulated annealing algorithm does not admit easy parallelization, because the state of the algorithm at iteration  $i$  is directly dependent on the state at iteration  $i - 1$ . Convergence to the same parameter values from multiple independent runs—separate instances of simulated annealing started from different randomly-selected states—can be used to suggest that the true global optimum has been found, but this technique does not improve the execution time of each annealing run. Another general approach for parallelizing simulated annealing is to perform multiple runs at different temperatures [39, 113]. High-temperature runs effectively search parameter space for locally-optimal regions, while low-temperature runs perform more fine-grained parameter optimization. At various times during the algorithm’s execution, the temperatures of two runs are swapped, allowing a low-temperature run to ‘jump’ to the state occupied by a high-temperature run and vice versa. Again however, it is not clear that this approach would significantly improve execution time compared to a single simulated annealing run.

Our runtime profiling data suggests that the likelihood calculation at each iteration is potentially time consuming, indicating that parallelization of this calculation could be an effective strategy for improving algorithmic efficiency. There are two dimensions in which the likelihood calculation can be made parallel: either the sequence data can be partitioned among processors, or the model’s parameters can be partitioned. In the first case, the likelihood scores for separate columns in the data matrix are calculated independently on different processors and then combined once all scores are available. This approach is especially appealing in the case of large genomic data sets like the bilaterian data analyzed above. When sequence length is very long, the size of the sequence data in memory can impact

performance by reducing cache efficiency. Partitioning these data into relatively large independent subsets might therefore show faster-than-linear speedup. Data-partitioning parallelism is probably less useful when sequence length is relatively small.

In the case of the mixed branch length model and other mixed models—such as the discrete gamma model of among-site rate variation—it is possible to partition sub-models among processors. In this case, likelihood scores for different sets of branch lengths are calculated for the entire data matrix independently; these scores are then trivially combined to obtain the total likelihood. The appeal of this approach is that model complexity has only a minimal impact on execution time: a simple model running on a single processor takes the same amount of time to calculate as a complex model running on multiple processors. In cases where the best-fit model is very complex, this approach should provide good speedup compared to a sequential algorithm. The drawback to this method is that algorithmic efficiency cannot be improved when the model is simple, no matter how many processors are available.

In order to provide an adaptive parallelization strategy useful both in situations where sequences are long and in situations where the evolutionary model is complex, we propose to implement both sequence-data partitioning and sub-model partitioning parallelism. User-tunable parameters will control the level of partitioning in each dimension, and we will perform extensive analyses to empirically determine optimal levels of partitioning for a number of data sets in order to guide the user's choice.

## 4.7 Conclusion

Simulated annealing is a powerful and general strategy for optimizing complex functions. Although simulated annealing has been used previously to address phylogenetic problems, no current implementations make full use of the power of the algorithm, and many are not useful for general phylogenetic reconstruction. The LVB algorithm proposed by Barker [7] implements maximum parsimony

optimization only and is not available for likelihood-based methods. The RAxML-SA method of Stamatakis [105] uses simulated annealing to find the optimal topology but relies on numerical optimization to estimate branch lengths and other model parameters, making it useless for optimizing complex mixed models. Finally, the method implemented by Salter and Pearl [97] is only useful for reconstructing rooted trees under the molecular clock assumption (an assumption typically violated by real sequence data), does not implement among-site rate variation models, and cannot be used to optimize model parameters other than branch lengths. Our software is the only complete implementation of a simulated annealing approach to phylogenetic reconstruction, and ours is the only implementation of the mixed branch length model in a maximum likelihood framework.

Even though our implementation has been successfully applied to moderate-sized phylogenetic problems, we have encountered computational difficulties attempting to analyze larger problems—either in terms of increased taxon sampling or longer sequences. Runtime profiling of our prototype software has identified a number of potential areas for code optimization. In addition, improvements to the simple simulated annealing algorithm we implemented have been suggested by other authors; these could be implemented to improve the efficiency of our implementation. Generalizing the simulated annealing algorithm to optimize model complexity in addition to topology and model parameters may also improve runtime performance, and parallelization strategies partitioning either data or model parameters among processors are also potentially useful. Future research will examine the effectiveness of each of the strategies outlined above for improving algorithmic efficiency and runtime performance.

## CHAPTER V

### IS THERE A STAR TREE PARADOX?

This chapter was originally published in the journal *Molecular Biology and Evolution* (vol. 25 no. 10, pp. 1819-1823, 2006). It was co-authored by Joseph W. Thornton, who assisted with experimental design and edited the manuscript.

#### 5.1 Introduction

Accurately characterizing statistical confidence in phylogenetic hypotheses is an important and long-standing challenge. Bayesian phylogenetics expresses confidence in terms of posterior probability—the probability that a tree or clade is true given the data, an evolutionary model, and prior probability distributions over model parameters [54]. There is growing concern that inferred posterior probabilities may be generally “overcredible” in a frequentist sense, leading to inflated confidence in uncertain relationships and a high rate of incorrect inferences, especially when the true tree has zero- or near-zero length internal branches, [12, 67, 112, 132].

Most software for Bayesian phylogenetic inference uses Markov Chain Monte Carlo (MCMC) techniques that sample only resolved trees; unresolved phylogenies are approached by examining very short internal branch lengths (typically  $10^{-6}$  substitutions/site), but the remaining “hole” in parameter space is not sampled. Lewis, Holder, and Holsinger ([67], LHH) suggested that when the true tree is unresolved, not sampling unresolved trees causes “disturbingly high” posterior probabilities to be inferred for one or another arbitrarily resolved tree, and this

problem gets worse as sequence length increases: “For large data sets, the phylogenetic uncertainty generated by the true polytomy manifests itself as unpredictability in the level of estimated posterior support for arbitrary resolutions of the polytomy, not as increased homogeneity of support for all possible resolutions.” This conclusion was based on a series of simulations using a four-taxon star tree with equal terminal branch lengths: when replicate sequences of length  $N = 1$  were analyzed, equal posterior probability was always inferred for each possible resolved tree, but when longer sequences ( $N = 100,000$ ) were analyzed, some replicates produced high support for one of the three resolved trees. Yang and Rannala ([132], YR) examined additional sequence lengths and also found that very short sequences ( $N = 20$ ) always produced roughly equal support for each resolved tree, but longer sequences ( $N = 200$  and  $N = 1000$ ) occasionally yielded high support for one of the three topologies. Both LHH and YR sketched theoretical arguments predicting that “posterior probabilities of particular resolutions of polytomous tree topologies will become more unpredictable with increasing sequence length” [67] and become completely unpredictable as  $N$  approaches infinity. If true, this “star tree paradox” is a real concern; it suggests that posterior probabilities on trees with short internal branches may regularly generate inflated confidence in incorrect or uncertain phylogenies, leading to frequent inferences of incorrect evolutionary relationships and increasingly pathological behavior as more data are analyzed.

The prediction that posterior probabilities would become more problematic as sequence length grows has not been directly tested, however. Both LHH and YR examined too few sequence lengths to establish a general trend, and neither explicitly examined how BMCMC methods would perform with infinite data. Further, neither study investigated whether at any sequence length high posterior probabilities are observed more often than they should be. Here we use simulation experiments to test the predictions associated with the star tree paradox.

## 5.2 Methods

Posterior probabilities were estimated using MrBayes v3.1 [95]. Four incrementally heated chains ( $temp = 0.2$ ) were run for 205,000 generations, with samples taken every 100 generations. The first 5000 generations were discarded as burnin. Prior probabilities were equal over all tree topologies and uniformly distributed on  $[0,10]$  for branch lengths. The shape parameter for gamma-distributed among-site rate variation was given a uniform prior on  $[0.05,50]$ , and the default prior was used for the transition/transversion ratio. The true evolutionary model was assumed.

Sequence alignments of length 1, 10, 100,  $10^3$ ,  $10^4$ ,  $10^5$ ,  $10^6$ , and  $10^7$  nucleotides were simulated on a four-taxon star tree. We simulated data using the JC69 substitution model and either equal terminal branches (0.5 or 0.05 substitutions/site) or Felsenstein-zone branch lengths with two long terminals (0.75) and two short terminals (0.05). Data were also simulated using the K80+g ( $\kappa = 10, \alpha = 0.5$ ) model and equal long terminal branches (0.5). We analyzed 1000 replicate alignments under each set of experimental conditions as described above. In addition, we analyzed 5000-nt alignments simulated using a 10-taxon star tree with long terminal branches (0.5), with posterior probabilities on clades summarized over trees using MrBayes. In each case, observed type I error rates were compared to maximum acceptable values using a one-sided  $t$  test.

To examine the accuracy of posterior probabilities with infinite data, ideal pseudo-datasets with no stochastic error were analyzed. We calculated the expected frequency of each character state pattern ( $f(x)$ ) under a four-taxon star phylogeny with either long (0.5) or short (0.05) terminal branch lengths and the JC69 substitution model. We modified the source code of MrBayes v3.1 to estimate posterior probabilities given this vector of state pattern frequencies. The per-site likelihood of tree  $t$  given any state pattern  $x$  is calculated by raising the probability of the pattern, given the tree, to the frequency with which that pattern is expected to occur:  $L(t|x) = P(x|t)^{f(x)}$ . The total per-site likelihood of the tree is the product of this partial likelihood over all possible state patterns. Each ideal dataset was

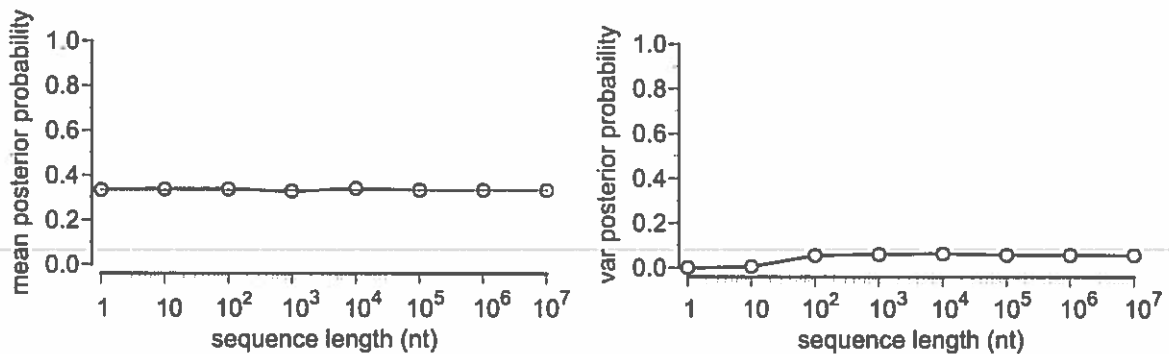


analyzed 100 times to account for variation in Markov chain sampling. We calculated the mean posterior probability over all analyses and assessed deviation from expected support of 1/3 for each possible resolved tree using a *t* test.

### 5.3 Results

If posterior probabilities become increasingly unpredictable as sequence length increases, then the variance in the posterior probability of a particular resolved tree over replicate datasets should increase as  $N$  grows. We tested this prediction by simulating alignments of various lengths on a four-taxon star tree using conditions similar to those examined by LHH and YR and estimating the posterior probability of a resolved phylogeny using MrBayes v3.1, which does not sample unresolved trees. We found that the mean posterior probability is always close to 1/3, and—after an initial increase—the variance remains stable with increasing sequence length (Figure 5.1). When  $N \leq 10$ , the variance in posterior probability is close to zero, because a resolved tree can only be supported by convergent substitutions on at least two branches; in very small datasets, such low-probability patterns usually do not occur at all. Once sequences are long enough for convergent patterns to appear, however, there is no increase in variance with the amount of data analyzed. The apparent increase in earlier studies was due to a failure to examine enough sequence lengths to distinguish between a long-term trend and the initial increase due to near-zero variance at extremely short sequence lengths.

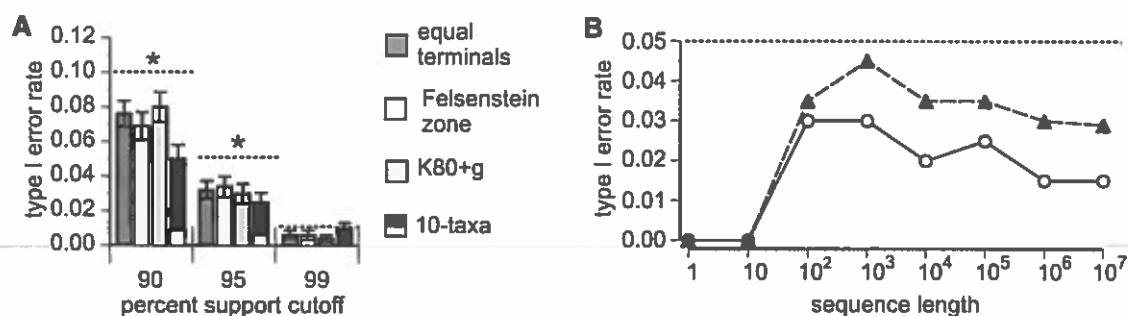
Both LHH and YR noticed that, when moderate or long sequences are simulated on a four-taxon star tree, “disturbingly high” posterior probabilities were occasionally observed, in contrast to very short sequences, for which posterior probabilities were always close to 1/3. The occasional presence of high posterior probabilities is not in itself reason for concern. Although a star tree is expected to generate equal frequencies of state patterns that support each of the three trees, stochastic variation with finite datasets causes pattern frequencies to deviate from expectation, leading to spurious support for one tree or another. Usually the stochastic deviation is small, but infrequently it will be larger. Unequal pattern



**FIGURE 5.1:** Variance in posterior probability of a resolved tree does not increase with increasing sequence length. The mean (left panel) and variance (right panel) in posterior probability of a particular resolved tree over replicate datasets is shown when data are generated using the JC69 substitution model and a four-taxon star tree with long terminal branch lengths (0.5 substitutions/site). Gray line indicates expected mean support of  $1/3$ .

frequencies also occur when the true tree is resolved, producing phylogenetic signal. The purpose of posterior probabilities is to help distinguish these possibilities by expressing the probability that some resolved tree is true given the data. If a method is to have any power to detect a resolved phylogeny when it is true, high posterior probabilities must occur occasionally when finite data are generated on the star tree. The crucial question is whether they occur more often than should, leading to a high rate of erroneous inferences—an issue not addressed by LHH or YR.

If the posterior probability of a tree accurately estimates the probability that the tree is the true tree (which it has been shown to do when the true tree is resolved, provided the model and priors are correctly specified [53, 132]), a hypothesis with posterior probability 0.95 should have a 0.05 chance of being false, and a group of hypotheses with posterior probability 0.95 should contain 5% incorrect trees. Though unconventional in a Bayesian framework, the use of a decision rule that accepts a phylogenetic hypothesis only if it has posterior probability  $> 0.95$  should therefore result in a long-run type I error rate  $< 0.05$  if posterior probabilities accurately measure the frequentist probability that a tree is correct. We tested whether use of current BMCMC implementations leads to high rates of type I error



**FIGURE 5.2:** Type I error rates based on posterior probability are conservative. **A)** The fraction of star trees incorrectly resolved at support cutoff values of 0.90, 0.95, and 0.99 posterior probability is shown. Sequences were 5000 nt long. Dotted line indicates maximum permissible error rate. Bars indicate standard error, with a significantly reduced error rate compared to the maximum permissible for each cutoff being indicated by an asterisk ( $\alpha = 0.01$ ). For 10-taxon trees, we calculated type I error rates when posterior probabilities were summarized on clades in two ways: 1) each clade is considered an independent hypothesis (white), and 2) a single resolved clade per replicate is considered a type I error (black). **B)** Type I error rates (based on a 0.95 posterior probability cutoff) are shown as sequence length increases when the true four-taxon star tree has long terminal branches (0.5 substitutions/site, filled triangles) or short terminals (0.05, open circles). Dotted line indicates maximally acceptable type I error rate.

by simulating replicate sequence alignments on various unresolved trees and observing the proportion of resolved trees with posterior probabilities greater than various cutoff values. Resolved trees with support greater than cutoffs of 0.90, 0.95, and 0.99 were considered type I errors [115], and observed error rates were compared to maximum error rates expected if posterior probabilities are accurate estimators that a tree is true (0.10, 0.05, and 0.01, respectively). We found that type I error rates were lower than the maximum acceptable for all sequence lengths and thresholds used, whether terminal branch lengths were equal or in the more challenging Felsenstein zone, and whether data were generated using simple or complex evolutionary models (Figure 5.2A).

To determine if type I error rates remain low when larger phylogenies are analyzed, we simulated data using a 10-taxon star tree with long terminal branches (0.5 substitutions/site), analyzed these data using MrBayes, and assessed the

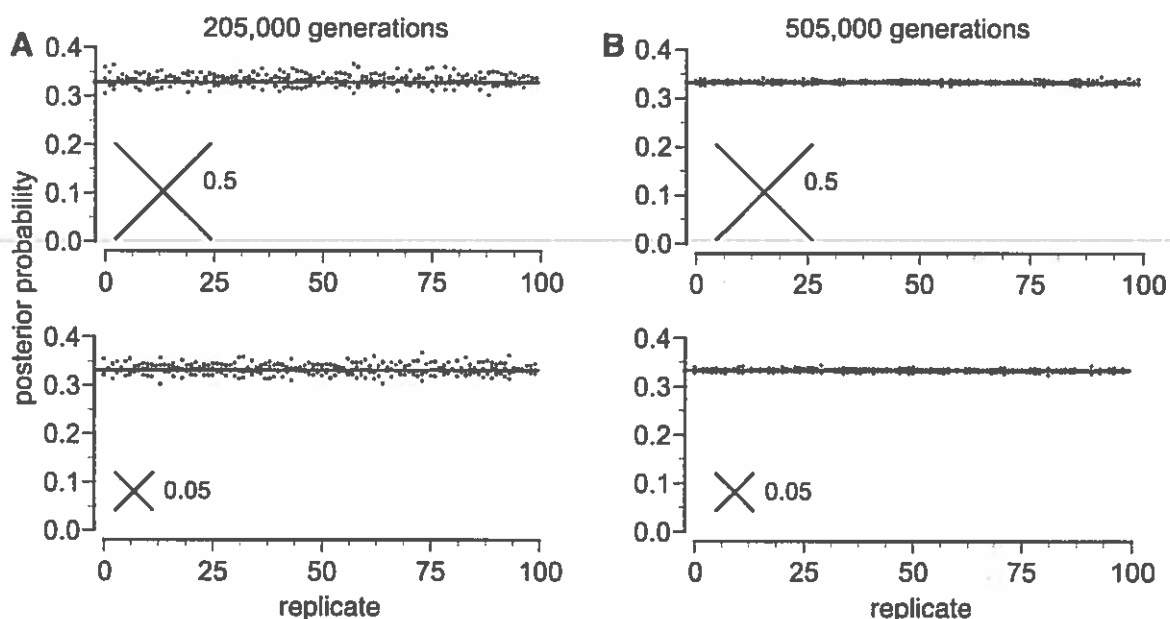
frequency with which incorrectly resolved phylogenetic hypotheses were strongly supported. First, we found that no fully-resolved trees were supported with posterior probability  $> 0.9$  when data were generated on a large star tree (not shown). Posterior probabilities are typically not reported on an entire tree, however; more commonly, posterior probabilities are reported on individual clades by summarizing over all sampled trees. Recent concerns have been raised that this practice may introduce a bias in posterior probability estimates, because the typical assumption of flat priors over trees places higher prior probability on clades with either few or many taxa [86]. We found that when posterior probabilities are summarized on individual clades—with each clade being considered an independent hypothesis—type I error rates are very low (Figure 5.2A). Even if a single strongly supported clade per replicate is considered a type I error—an extremely conservative measure of type I error rate—the rate of erroneous inferences is still less than maximally acceptable at all posterior probability cutoffs examined. Taken together, the results of our type I error analyses indicate that current BMCMC implementations do not frequently produce excessive confidence in falsely-resolved hypotheses when data are generated on a star tree, even though the true phylogeny is never explicitly considered.

Contrary to the prediction that posterior probabilities would become increasingly unreliable as sequence length increases, we found that type I error rates fall with increasing sequence length (Figure 5.2B). As observed with the variance in posterior probabilities, the trend in type I error with sequence length is not monotonic. With extremely short sequences ( $N \leq 10$ ), the error rate is close to zero, presumably due to the lack of any convergent state patterns. The rate of type I error then increases with sequence length as convergent patterns begin to occur, peaking at moderately short lengths ( $N = 100$  to  $1000$ ) and then falling as sampling error becomes less important with longer sequences. Even when type I error rates are at their maximum, posterior probabilities never produce strong support for incorrectly resolved phylogenies more often than they should.

LHH and YR suggested that, as sequences generated on a star tree approach infinite length, we would like the inferred posterior probability of each possible

resolved tree to become equal; however they predicted that these posterior probabilities would become “completely unpredictable” over replicates. We tested this prediction by analyzing pseudo-datasets that possess the same characteristics as infinite data. With infinitely long sequences, the frequency of each possible character state pattern equals the expected frequency, and the variance in pattern frequencies among datasets is zero. To determine the posterior probabilities that would be inferred if infinite data were available, we generated pseudo-infinite datasets that do not deviate from expected character state pattern frequencies and estimated posterior probabilities from these data by BMCMC without explicitly sampling the true unresolved tree. Specifically, we calculated the per-site likelihood of an infinite dataset by calculating the expected pattern frequencies given the simulation conditions and modifying MrBayes to infer posterior probabilities given a list of patterns and their associated frequencies. When these ideal data were repeatedly analyzed, we observed a mean posterior probability of 0.333 for each possible tree (Figure 5.3A) and very little scatter about the mean ( $\sigma^2 = 1.2 \times 10^{-4}$ ). From 200 replicates—100 with long terminal branches (0.5) and 100 with short terminals (0.05)—the maximum posterior probability observed for any resolved tree was 0.37. The small amount of variation observed among replicates appears to be due to stochastic error in MCMC sampling: when longer runs are executed, posterior probabilities are even closer to 1/3 (Figure 5.3B,  $\sigma^2 = 1.06 \times 10^{-5}$ , maximum posterior probability=0.34). These results indicate that posterior probabilities do produce equal support for all resolved trees in the infinite case ( $P = 0.98$ ), which is the desired result. Analysis of ideal datasets does not indicate what will happen when very large datasets with some stochastic error are analyzed, but it does show that when infinite data are generated on a star tree, posterior probabilities are predictable, equally supporting each possible resolved tree.

LHH and YR both used a coin-flipping analogy to support their contention that posterior probabilities become increasingly unpredictable as sequence length grows. YR demonstrated that when the null hypothesis is true (i.e., the coin is fair), the expected frequency distribution of posterior probabilities for each “resolved” hypothesis (that the coin is biased one way or the other) is uniform. Although the



**FIGURE 5.3:** Star-tree generated data with ideal character state pattern frequencies produce equal posterior probability for each possible resolved tree. The inferred posterior probability of each resolved four-taxon tree (black dots) is shown for independent Bayesian analyses of an ideal dataset with no stochastic variation. In the top panel, the true tree has four long terminal branches (0.5 substitutions/site); in the lower panel, the true tree has short terminals (0.05). Solid lines indicate theoretically correct inference of 1/3 support for each tree. MCMC analyses were run for either 205,000 generations (A) or 505,000 generations (B).

distribution of posterior probabilities on phylogenetic trees is unknown, LHH and YR presented the coin-flipping result as evidence of pathological behavior when the null hypothesis is true but is not explicitly examined. In fact, this behavior is reassuring. The uniform frequency distribution implies that data leading to an inferred posterior probability  $\geq 0.95$  on incorrect trees will be observed at most 5% of the time, data leading to a posterior probability  $\geq 0.90$  will be observed 10% of the time, etc.—precisely the behavior expected if posterior probabilities accurately reflect the probability that the hypothesis is true. For the same reason, a uniform distribution is also observed for frequentist  $P$ -values whenever the null hypothesis is true. The initially appealing intuition that posterior probabilities should converge on equal support for each resolved hypothesis is correct only when data are truly

infinite and precisely match the null expectation; in this case the frequentist  $P$ -value will be 1.0. If posterior probabilities calculated from finite data were instead concentrated around  $1/T$  (where  $T$  is the number of possible resolved hypotheses, a situation similar to having  $P$ -values biased toward 1.0 when the null hypothesis is true), we would always infer low posterior probabilities for resolved trees—not only when the null hypothesis is true but also when the true hypothesis is resolved—leading to reduced statistical power to resolve difficult phylogenies.

## 5.4 Discussion

The implication of our results is that there is no star tree paradox. Even when trees contain zero- or near-zero length internal branches, posterior probabilities behave as an appropriate statistical estimator should, providing near-equal support for all possible resolved trees with infinite sequence length and producing strong support for incorrect trees very infrequently when finite data are analyzed. The fact that unresolved trees are not explicitly evaluated has no apparent effect on the accuracy of posterior probability as a measure of statistical confidence. Furthermore, we have shown that evidence previously presented in favor of the star tree paradox has been erroneously interpreted. The occasional support in favor of a falsely-resolved phylogeny observed by LHH and YR is the expected result of stochastic error, and the convergence of posterior probabilities to the uniform distribution is a desirable property of a statistical estimator, producing a reasonable balance between power to resolve difficult problems with strong support and a low rate of false inferences. Our results do not imply that posterior probabilities will never be inflated; previous studies have shown that posterior probabilities can be unreliable when either the evolutionary model [53, 112] or prior assumptions about model parameters [132] are incorrectly specified. That existing methods do not sample unresolved trees, however, does not inflate posterior probabilities inferred by MCMC.

---

## CHAPTER VI

# EFFECTS OF PRIOR BRANCH LENGTH UNCERTAINTY ON BAYESIAN POSTERIOR PROBABILITIES FOR PHYLOGENETIC HYPOTHESES

### 6.1 Introduction

One of the central goals of statistics is to help us express how certain we are when we make an inference based on evidence. Phylogenies provide the framework for all valid comparative biology, so reliable measures of statistical confidence in evolutionary trees have long been sought. The most common method is nonparametric bootstrapping [23], a resampling procedure that has been shown to be conservatively biased [44, 103]. Bayesian phylogenetics [54, 92] expresses statistical support in terms of posterior probability, which is the probability that a tree is correct given the data, a model of the evolutionary process, and prior probability distributions over trees and model parameters [51]. Bayesian methods are widely applied in phylogenetics, having resolved difficult and long-standing problems with strong support [58, 79]. Bayesian Markov Chain Monte Carlo (BMCMC) algorithms allow large phylogenies to be efficiently estimated using complex evolutionary models, and posterior probabilities provide an intuitively meaningful measure of statistical confidence.



Concerns have been raised, however, that posterior probabilities may not accurately estimate statistical confidence. Although a number of studies support the general reliability of the Bayesian approach—while cautioning that results can be sensitive to model misspecification [3, 10, 20, 53, 65, 124]—others have directly challenged these findings, concluding that posterior probabilities are regularly “overcredible” and produce high rates of false inferences, even when the correct model is used [12, 17, 67, 74, 102, 112, 117]. Whether posterior probabilities are inflated always or only under some conditions—and why—remains an open question. As a result, the confidence we should have in phylogenies inferred using Bayesian techniques is uncertain.

Much of the controversy surrounding the reliability of posterior probabilities in phylogenetics may actually stem from a misunderstanding of what posterior probabilities on trees actually mean. Early theoretical arguments suggested that posterior probabilities calculated using uninformative priors should be equivalent to bootstrap proportions [19]. Comparative studies generally contradicted this prediction, finding instead that posterior probabilities are typically higher than bootstrap confidence [12, 17, 20]. In fact, posterior probabilities and bootstrap proportions are two different approaches to estimating statistical confidence and are not expected to be equivalent [3]. Bootstrap proportions attempt to estimate how often a given phylogeny would be recovered if replicate data sets could be sampled from the same process that generated the original data. In contrast, posterior probabilities measure the degree of support for a given phylogeny from the data set actually sampled, conditional on the model of evolution and prior assumptions about model parameters.

Because Bayes' Theorem states that the posterior probability of a hypothesis is the probability that the hypothesis is correct—given the data, a model of the data-generating process, and prior probability distributions over model parameters—numerous studies have compared posterior probabilities on phylogenies to the proportion of inferred trees that are correct [20, 53, 74, 102, 117, 124, 132]. However, there is no a priori expectation that posterior probabilities should necessarily indicate the proportion of inferences that are correct, because 1)

posterior probabilities are calculated directly from the data at hand and do not require replication, whereas calculating the percent of correctly inferred trees necessarily requires that a large number of inferences be made from different data sets, and 2) posterior probabilities are conditioned on prior assumptions, while the proportion of correct trees is not.

There are some specific conditions, however, in which the average posterior probability of a group of inferences is expected to equal the proportion of correct inferences in the group. The first condition necessary for this relationship to hold is replication. Although evolutionary history happened once and cannot be replicated, computer simulations allow us to generate multiple replicate data sets from any conceivable set of evolutionary conditions, so the proportion of correct inferences can be calculated. Under simulation conditions, when the chance of choosing each set of evolutionary parameter values to generate data is known in advance and used as prior information in a Bayesian analysis (i.e. the true priors are used), the average posterior probability of a group of inferences is equivalent to the proportion of those inferences that are correct. This follows directly from Bayes' Theorem and has been empirically shown in the case of phylogenetic inference when branch lengths are exponentially distributed [53, 132].

These results suggest that, were the actual values of nuisance parameters known in advance, the average posterior probability of a group of hypotheses would equal the proportion of correct hypotheses in the group. In real analyses, the values of nuisance parameters are never known in advance. Bayesian analysis incorporates this prior uncertainty by integrating over many parameter values, conditioned on prior beliefs about the probability of potential values for each model parameter. Unfortunately, little is known about the effects of integrating over uncertainty using different prior assumptions on resulting posterior probabilities for phylogenetic hypotheses, and the robustness of the perfect correspondence between posterior probability and proportion correct to prior uncertainty about the values of nuisance parameters has not been thoroughly examined.

Yang and Rannala [132] simulated sequence data on rooted three-taxon trees with branch lengths drawn from exponential distributions and compared the

posterior probability of a group of trees to the proportion correct when different exponential priors were assumed for terminal and internal branch lengths. When the the same distributions used to generate data were also used as priors for Bayesian analysis, the average posterior probability of a group of trees was the same as the proportion of trees that were correct. However, when the mean of the prior distribution on the internal branch length was greater than the actual mean, posterior probabilities were higher than the proportion of correct trees; when the mean of the internal branch length prior was less than the true mean, posterior probabilities were lower. Additionally, when the land plant data of [58] were analyzed using exponential priors with very small means on internal branch lengths, many of the inferred clades exhibited reduced posterior probabilities.

Yang and Rannala's experiments established that the choice of branch length priors can affect posterior probabilities, and that the perfect correspondence between posterior probability and proportion correct is not necessarily robust to prior uncertainty when certain priors are used, but several important questions remain unresolved. First, the simulations employed by previous authors [53, 132] represent a peculiar situation in which phylogenetic trees and branch lengths are generated by a stochastic process. Real evolutionary history is not generated stochastically but follows a single historically correct tree. How different prior assumptions affect posterior probabilities when there is a single correct tree with fixed branch lengths is unknown. Second, although Yang and Rannala examined various priors for the internal branch, a separate prior—always the actual distribution used to simulate data—was independently assigned to terminal branches. In most real analyses, however, a single prior distribution is applied to all branches on the tree. How different branch length priors applied across the entire tree affect posterior probabilities is unknown. Third, it has been common to use a uniform prior distribution with a large upper bound on branch lengths to represent prior ignorance about this parameter; because such a prior will usually overestimate mean branch lengths, Yang and Rannala predicted that flat priors would produce excessively high posterior probabilities on trees. Yang and Rannala recommended against such priors and suggested an exponential prior with very small mean.

Whether flat branch length priors actually produce high posterior probabilities has, however, not been tested. Finally, Yang and Rannala considered only a single pattern of branch lengths; different branch length patterns might interact with prior assumptions to produce different effects.

Here we address these questions by examining the effects of integrating over branch length uncertainty using various prior distributions on posterior probabilities calculated for phylogenetic trees. We show that branch length uncertainty can affect posterior probabilities across a range of phylogenetic problems and when various prior distributions are used. Although posterior probabilities can be relatively stable when different diffuse priors are assumed (including flat priors with various upper bounds and exponential priors with moderate to large mean values), using an exponential prior with very small mean across the entire tree produces more extreme posterior probabilities, resulting in a higher frequency of incorrect inferences with strong support. Additionally, posterior probabilities inferred using any of the typical prior distributions can differ significantly from those calculated when the actual branch lengths are known in advance. When branch lengths are not known with certainty, the pattern of branch lengths on the true tree has a strong effect on posterior probabilities, sometimes causing them to deviate significantly from the posterior probabilities that would be inferred if the true branch lengths were known. Some patterns push posterior probabilities upward, while others push posterior probabilities downward. We conclude that prior uncertainty about branch lengths—and potentially other parameters as well—interacts with sequence length, the pattern of branch lengths on the tree, and the prior distributions assumed for the analysis to produce a complex effect on posterior probabilities with potentially significant consequences for phylogenetic practice.

## 6.2 Methods

### 6.2.1 Bayesian Analyses

Posterior probabilities were estimated by MCMC using MrBayes v3.1 [95]. Four incrementally heated chains ( $temp = 0.2$ ) were run until the average standard deviation in posterior probability estimates across two independent BMCMC runs was  $< 0.01$ , indicating that chains had run long enough to converge on the posterior distribution. Prior distributions on branch lengths were either uniform on  $(0, M]$  ( $M = 1, 5, 10$ , or  $100$ ) or exponential with  $\mu = 10^{-5}, 10^{-3}, 0.01, 0.1$ , or  $1.0$ . (Note that MrBayes uses  $1/\mu$  to parameterize the exponential distribution, so  $\mu = 10^{-5}$  corresponds to a parameter value of  $100,000$ ) The true model was used for all BMCMC analyses.

In addition to BMCMC analyses, we conducted Bayesian analyses using an empirical Bayes approach that places prior probability 1.0 on the maximum likelihood branch lengths calculated for each possible tree topology. We calculated the maximized likelihood of each tree using PAUP\* v4.0b10 [114] to optimize branch lengths. Posterior probabilities were then calculated directly from Bayes' Theorem using these maximized likelihood values.

### 6.2.2 Accuracy of BMCMC

We compared posterior probabilities estimated by BMCMC to Bayesian posterior probabilities calculated using Bayes' Theorem for a number of four-taxon problems using both linear regression and the  $\chi^2$  test, with posterior probabilities binned every 0.01, combining adjacent bins to assure each bin had  $> 5$  elements. One thousand alignments of 10,000 characters were simulated on the  $((AB),(CD))$  tree using either the JC69 or K80+G ( $\kappa = 10, \alpha = 2.0$ ) model. We examined two types of branch length combinations: 1) equal terminal lengths (0.5 substitutions/site) with internal branch lengths of 0.0, 0.01, and 0.03, and 2) Felsenstein zone trees with and internal branch length of 0.01, long terminals (0.75) leading to nonsister taxa A and C, and short terminals (0.05) leading to B and D.

We calculated the true posterior probability using Bayes' Theorem:

$$P(t_i|X) = \frac{\int_v \int_\kappa \int_\alpha P(X|t_i, v, \kappa, \alpha) P(t_i, v, \kappa, \alpha)}{\sum_j \int_v \int_\kappa \int_\alpha P(X|t_j, v, \kappa, \alpha) P(t_j, v, \kappa, \alpha)}$$

where  $v$  is the set of branch lengths,  $\kappa$  is the transition:transversion parameter, and  $\alpha$  is the gamma distribution shape parameter.  $P(X|t_j, v, \kappa, \alpha)$  is the probability of the data given specified parameter values, and  $P(t_j, v, \kappa, \alpha)$  is the prior probability of the parameter values. We assumed uniformly distributed priors for both BMCMC and Bayes' Theorem calculations; when prior probabilities are uniformly distributed, the posterior probability reduces to the likelihood of tree  $t_i$  to the sum of the likelihoods of all trees, with model parameters being integrated out. We numerically estimated the integral  $\int_v \int_\kappa \int_\alpha P(X|t_j, v, \kappa, \alpha)$  for each tree using a rectangular approximation, calculating likelihoods for each set of parameter values using PAML v3.14 [130]. Branch lengths increased in steps of 0.001 on the interval [0.0,0.1] for the internal branch and  $\pm 0.2$  away from the true value for the terminal branches.  $\kappa$  was sampled on the interval [8,12], and  $\alpha$  was sampled on [0,4] in steps of 0.4. The resulting samples from the likelihood surface were used to estimate the integrated likelihood for each possible tree. We validated these numerical posterior probability estimates by comparing estimates from 30 randomly-selected datasets to posterior probabilities calculated using wider intervals and more thorough sampling within intervals; we found the narrower intervals and sparser sampling to be highly accurate (not shown).

### 6.2.3 Simulations

We performed "Bayesian simulations" [53, 132] using a four-taxon phylogeny with fixed branch lengths (terminals 0.5 substitutions/site, internal 0.01). Five-hundred replicates with sequence lengths 100, 1000, and 10,000 nucleotides were simulated using the JC69 model and a randomly selected topology for each replicate. Sequences were analyzed by BMCMC using the uniform and exponential branch length priors described above as well as the true point priors on branch

lengths, which place prior probability 1.0 on the branch lengths actually used to simulate data (0.5 for terminal and 0.01 for internal branches). Additionally, we calculated posterior probabilities using an empirical Bayes approach that places prior probability 1.0 on the maximum likelihood branch lengths calculated for each tree.

To assess the effects of different branch length patterns on inferred posterior probabilities, we simulated 500 replicate data sets of either 100 or 10,000 nucleotides on four-taxon topologies with internal branch length 0.01 and six different terminal branch length combinations: 1) all short branches (0.01), 2) one long branch (0.75) and three short, 3) three long and one short, 4) all long branches, 5) inverse-Felsenstein zone lengths, with two long sister branches and two short sister branches, and 6) Felsenstein zone branch lengths, with two long nonsister branches and two short nonsister branches. Sequences were analyzed by BMCMC using either a uniform branch length prior ( $U(0, 10)$ ), an exponential prior with  $\mu = 0.1$ , or a small-mean exponential prior with  $\mu = 10^{-5}$ . In addition, we analyzed data sets using the empirical Bayes prior that places prior probability 1.0 on the maximum likelihood branch lengths for each tree.

To assess the impact of typical branch length prior assumptions on clade probabilities under realistic conditions, we simulated 100 data sets of various sequence lengths (1000–50,000 nt) using parameter values drawn from an analysis of real sequence data [79]. To maintain computational tractability, we analyzed a 30-taxon subset in which very closely related taxa were represented by a single species. We estimated the tree topology, branch lengths, and parameters of the GTR+I+G model by BMCMC using the original nucleotide data and then used these conditions to simulate replicate data sets. Posterior probabilities were estimated by BMCMC using the true model, with priors on transition model parameters fixed to their true values and branch lengths assumed to be either uniformly distributed on  $(0, 10]$  or exponentially distributed with  $\mu = 0.1$  or  $\mu = 10^{-5}$ . The first 10,000 generations were discarded as burnin, and analyses were terminated when the average standard deviation in clade probabilities between two independent runs dropped below 0.01.

We also performed Bayesian simulations using a 10-taxon problem with increasing sequence length (25–1000 nucleotides) and the JC69 model. For each of 500 replicate data sets at each sequence length, the tree topology was selected at random and each branch length drawn from an exponential distribution with  $\mu = 0.15$ . BMCMC analyses were performed using three different branch length priors:  $U(0, 10)$ ,  $exp(10^{-5})$ , and  $exp(0.15)$ .

### 6.2.4 Comparing Posterior Probabilities

In order to assess the effects of integrating over branch length uncertainty using various prior distributions on posterior probabilities, we determined whether posterior probabilities calculated using different branch length priors matched those that would be inferred if the true branch lengths used to simulate data were known in advance. For each branch length prior, we collected posterior probabilities on trees into 10 equally-sized bins and compared the average posterior probability of each bin to the proportion of correct trees in the bin; these values should be equal if nuisance parameters are known with certainty [53, 132]. Bins with fewer than 20 trees were excluded to avoid stochastic error in estimating the proportion correct. Additionally, we examined the false-positive inference rate incurred using each branch length prior. We considered inferred trees with  $\geq 0.95$  posterior probability as strongly supported and calculated the proportion of replicate data sets producing incorrect inferences with strong support using each prior distribution.

## 6.3 Results

### 6.3.1 Accuracy of BMCMC

We used simulated data under a variety of conditions to evaluate the potential effects of branch length uncertainty on posterior probabilities. MrBayes—the most popular Bayesian phylogenetics software package—was used to estimate posterior probabilities by MCMC. Although MCMC should theoretically estimate posterior

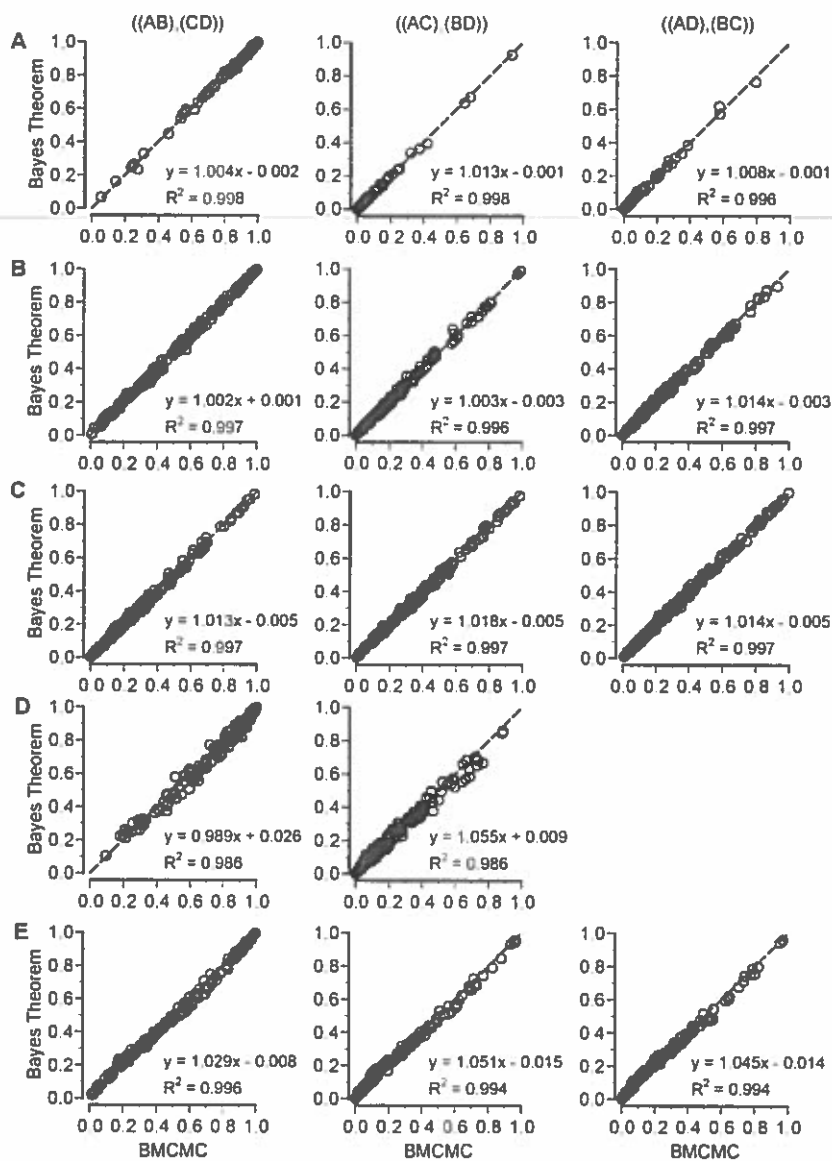


probabilities with high accuracy, the accuracy of MrBayes has not been determined experimentally. Therefore, to ensure that software errors did not undermine the validity of our results, we first verified that posterior probabilities estimated using MrBayes were equivalent to those calculated directly from Bayes' Theorem using numerical integration over branch lengths and other model parameters.

To determine if current BMCMC procedures correctly estimate posterior probabilities defined by Bayes' Theorem, we simulated sequence evolution under a variety of conditions on four-taxon trees and compared the posterior probability of each possible tree estimated by BMCMC (BMCMC-PP) to the posterior probability defined by Bayes' Theorem, which we calculated directly using numerical estimation of likelihoods integrated over branch lengths and other parameters (see Methods). We found that BMCMC-PPs are very accurate estimators of the true Bayesian posterior probability (Figure 6.1). When terminal branch lengths were equal (Figure 6.1A-C), BMCMC-PPs tightly fit the ideal line expected if they are the same as the true values ( $\chi^2 P > 0.995$ ). Even when the true tree was unresolved, BMCMC-PPs estimated without explicitly sampling zero-length internal branches were equivalent to numerical estimates calculated including unresolved trees ( $\chi^2 P = 0.997$ , Figure 6.1C). There was a slight reduction in accuracy when the true tree was a challenging Felsenstein-zone problem (Figure 6.1D), indicated by an increased scatter around the ideal regression line ( $\chi^2 P > 0.720$ ), but this reduction was minor and did not introduce bias. BMCMC-PPs were also highly accurate when the evolutionary model was more complex ( $\chi^2 P > 0.803$ , Figure 6.1E). We find no evidence for an intrinsic bias or overcredibility associated with the use of BMCMC to estimate posterior probabilities on phylogenies, at least for small trees. This allows us to rule out MCMC error as a potential factor in the four-taxon experiments described below.

### 6.3.2 Uncertainty Affects Posterior Probabilities

Bayesian phylogenetic analysis requires the specification of prior probability distributions over trees and model parameters—including branch lengths. Although



**FIGURE 6.1:** BMCMC accurately estimates Bayesian posterior probabilities. The posterior probability of each possible tree estimated by MCMC is plotted against the true Bayesian posterior probability calculated directly from Bayes' Theorem (see Materials and Methods). Sequences in A–D were simulated under a simple JC69 model; A–C had equal terminal branch lengths and internal branches of 0.03 (A), 0.01 (B), or 0.0 (C) substitutions/site, while panel D shows results using Felsenstein zone branch lengths and an internal branch of 0.01. Sequences in E were simulated using a more complex K80+G model with equal terminal branches and an internal branch length of 0.01.

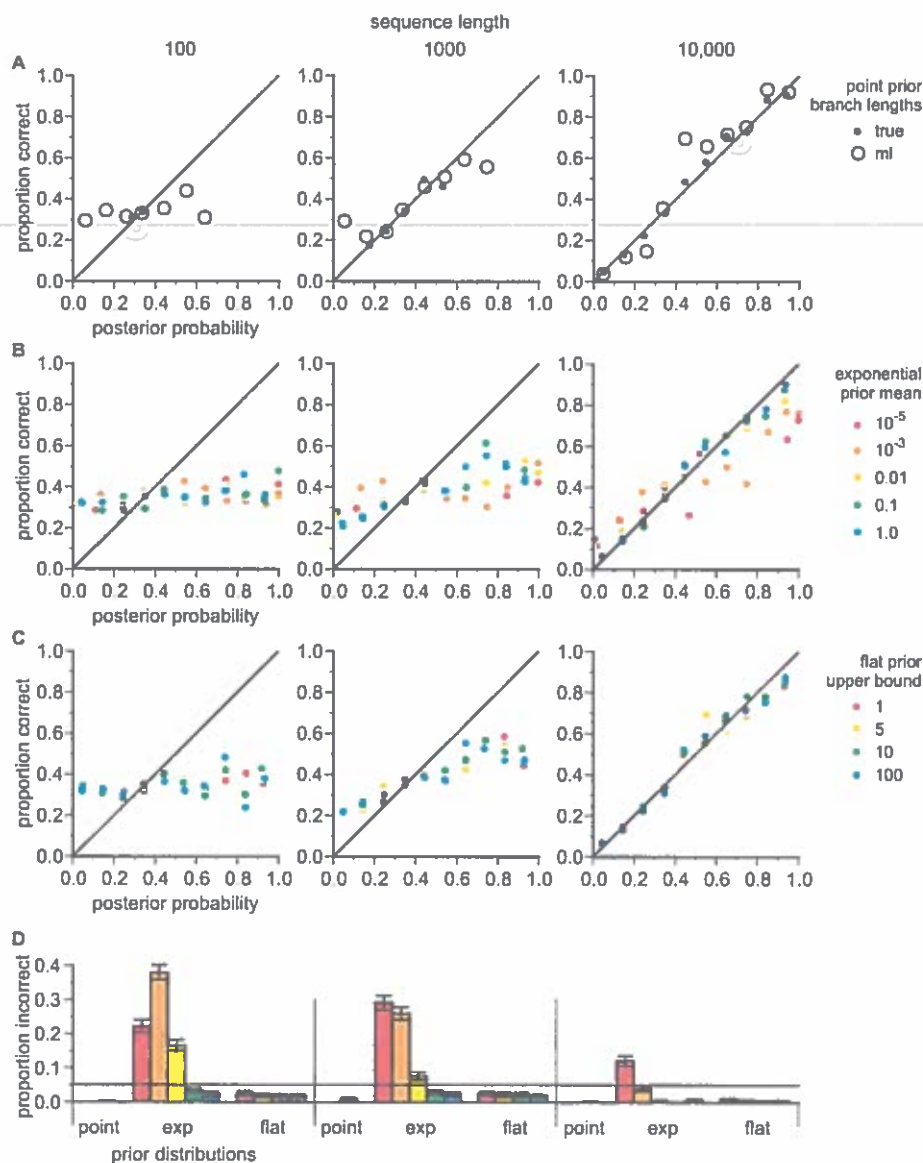
ultimately up to the researcher, the appropriate choice of prior distributions is not uncontroversial. It is intuitively appealing to assume 'flat' or 'uninformative' priors over branch lengths and other parameters to reflect prior ignorance concerning these values; this allows the likelihood function to be directly reflected by the posterior probability distribution without introducing any strong prior information that could skew the posterior distribution away from that produced directly from the data at hand. However, the assumption of flat priors is not itself unproblematic, because priors that are flat when parameters are scaled one way may be highly skewed if a different scale is used [24, 133]. If we have no prior information about the value of a parameter  $\theta$ , then we are equally ignorant about the potential values of  $\theta^2$ , but there is no prior distribution that is flat over both  $\theta$  and  $\theta^2$ . Completely uninformative priors are therefore impossible. Additionally, when a parameter is unbounded—as when branch lengths can range from zero to infinity—the prior distribution must be truncated, and it has been suggested that the choice of where to truncate the distribution can affect which values are contained in resulting credible intervals [24]. These concerns have led some researchers to eschew flat priors over branch lengths in favor of exponential distributions; in fact, the default branch length prior in MrBayes v3.1 is an exponential distribution with  $\mu = 0.1$ . Recently, it has been suggested that when the mean of the prior distribution on internal branch lengths is greater than the actual length, resulting posterior probabilities can be skewed upward [132]; this led the authors to recommend using an exponential prior with very small mean to avoid overcredibility.

To determine the potential effects of different branch length priors on posterior probabilities estimated for phylogenetic trees, we first simulated data of various lengths on a four-taxon phylogeny with equal terminal branch lengths (0.5 substitutions/site) and a short internal branch (0.01). Data were analyzed using flat priors with various upper bounds and exponential priors with different means. To isolate the effects of integrating over prior uncertainty, we analyzed data using two point prior distributions that avoid integration by fixing branch lengths. First, we used the true point prior distribution that places prior probability 1.0 on the actual branch lengths used to simulate the data, producing posterior probabilities given

perfect prior knowledge. Since perfect prior knowledge is never actually available, we additionally employed an empirical Bayes approach that places prior probability 1.0 on the maximum likelihood branch length estimates obtained from the data. This approach avoids integration over multiple branch lengths but does not rely on knowing the correct lengths in advance.

We compared posterior probabilities estimated using each prior distribution by collecting inferred posterior probabilities from replicate data sets into 10 equally-sized bins and plotting the mean posterior probability for each bin against the proportion of correct trees in that bin (Figure 6.2A-C). When the true values of nuisance parameters are known in advance, Bayes' Theorem suggests that these two values will be equivalent. Previous studies have empirically confirmed that when branch lengths are stochastically generated from an exponential prior distribution, and the same prior is used in a Bayesian analysis, the average posterior probability of a group of trees is equivalent to the proportion of those trees that are correct [53, 132]; our results show that this correspondence between average posterior probability and proportion correct also holds when branch lengths are fixed over replicate data sets (Figure 6.2A).

When branch lengths are unknown, they can either be estimated—by maximum likelihood for example—or integrated over using a more diffuse prior distribution. When branch length uncertainty was eliminated using an empirical Bayes approach that places prior probability 1.0 on the maximum likelihood branch lengths, the average posterior probability of a group of trees was higher than the proportion of those trees that were correct when sequences were short (100 nt), but longer sequences produced posterior probabilities that more closely matched the proportion of correct inferences (Figure 6.2A). When branch length uncertainty was incorporated by integrating over multiple values, the specific prior used affected the resulting posterior probabilities (Figure 6.2B,C). While all priors examined produced average posterior probabilities greater than the proportion of trees that were correct when sequences were of short or moderate length (100–1000 nt), long sequences (10,000 nt) resulted in average posterior probabilities that matched the proportion of correct inferences when diffuse priors were employed. For example,



**FIGURE 6.2:** Branch length uncertainty affects posterior probabilities. (A-C) The average posterior probability of a group of trees is plotted against the proportion of correct trees in the group for various sequence lengths (100–10,000 nt) and branch length prior distributions. Panel A shows results using point priors placing prior probability 1.0 on either the true branch lengths used to simulate data (black dots) or the maximum likelihood branch length estimates (open circles); panel B shows results from exponential priors with various means, and C shows results using flat priors with various upper bounds. (D) The proportion of replicate data sets giving strong support ( $\geq 0.95$  posterior probability) for incorrect trees is reported for each prior distribution and sequence length. Colors are the same as in panels A-C; bars indicate standard error, and horizontal line indicates an error rate of 0.05.

when exponential branch length priors were used, a larger mean on the exponential distribution (0.1–1.0 substitutions/site) produced posterior probabilities that closely matched the proportion of correct trees, while small-mean exponential distributions ( $10^{-5}$ –0.01) produced average posterior probabilities greater than the proportion of correct trees (Figure 6.2B). Flat branch length priors with upper bounds from 1–100 substitutions/site all produced similar posterior probabilities that closely matched the proportion of correct trees when sequences were long (Figure 6.2C).

To further explore the effects of branch length uncertainty on posterior probabilities, we calculated the proportion of false inferences with strong support ( $\geq 0.95$  posterior probability) produced at each sequence length by each prior distribution (Figure 6.2D). When the true branch lengths were known in advance, almost no false inferences were made. Similarly, estimating branch lengths using the empirical Bayes approach produced extremely low false inference rates. When branch length uncertainty was integrated over using exponential or flat prior distributions, the more diffuse distributions (flat priors with various upper bounds and exponential distributions with larger means) produced low false inference rates ( $< 0.05$ ), while small-mean exponential distributions produced excessive rates of false inferences with high posterior probability. In general, the rate of false inferences with strong support was higher when sequences were short and posterior probabilities were greater than the proportion of correct trees; longer sequences produced lower rates of false inferences.

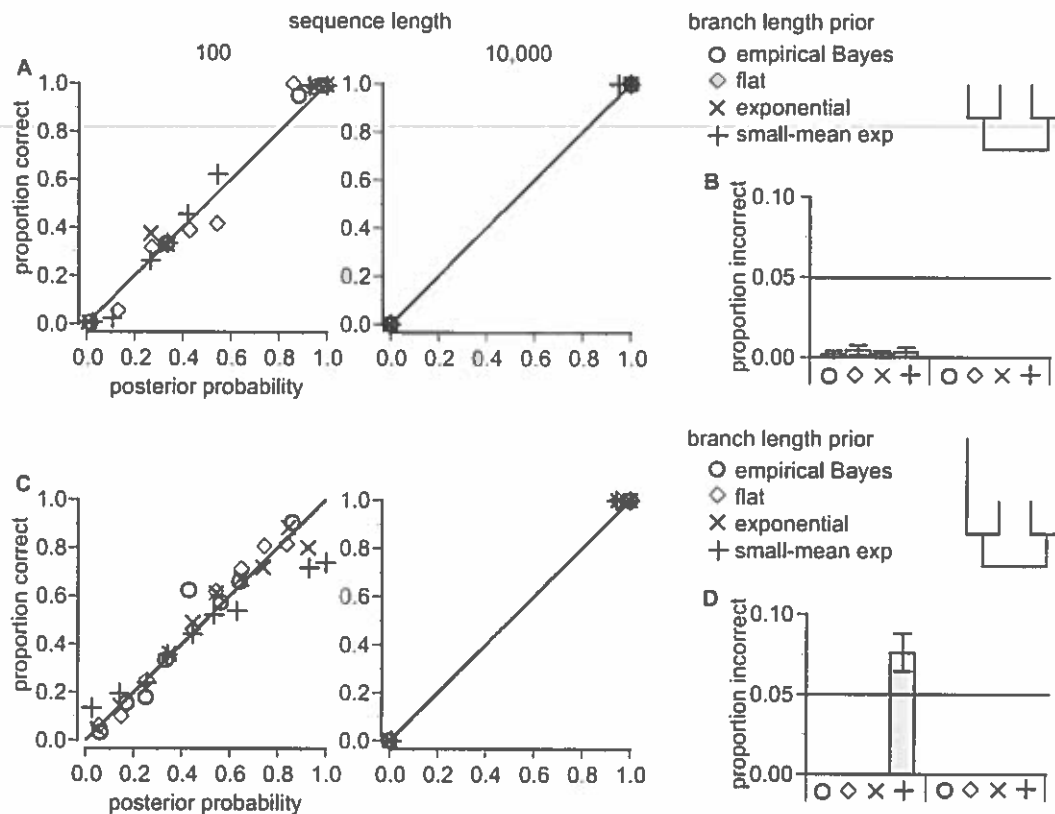
These results confirm that, given perfect prior knowledge about the values of nuisance parameters, the average posterior probability of a group of inferences is equivalent to the proportion of those inferences that are correct. Uncertainty about the values of nuisance parameters such as branch lengths can disrupt this relationship—causing the average posterior probability of a group of trees to be higher than the proportion of correct trees in the group—when sequences are not long enough to provide highly precise parameter estimates. Under the conditions examined, using either an empirical Bayes approach that places prior probability 1.0 on the maximum likelihood branch lengths or integrating over uncertainty using diffuse prior distributions produced posterior probabilities that more closely

matched the proportion of correct trees and resulted in lower rates of false inferences with strong support than using exponential priors with small means.

### 6.3.3 The Effects of Branch Length Uncertainty on Posterior Probabilities Are Determined by the Pattern of Branch Lengths on the True Tree

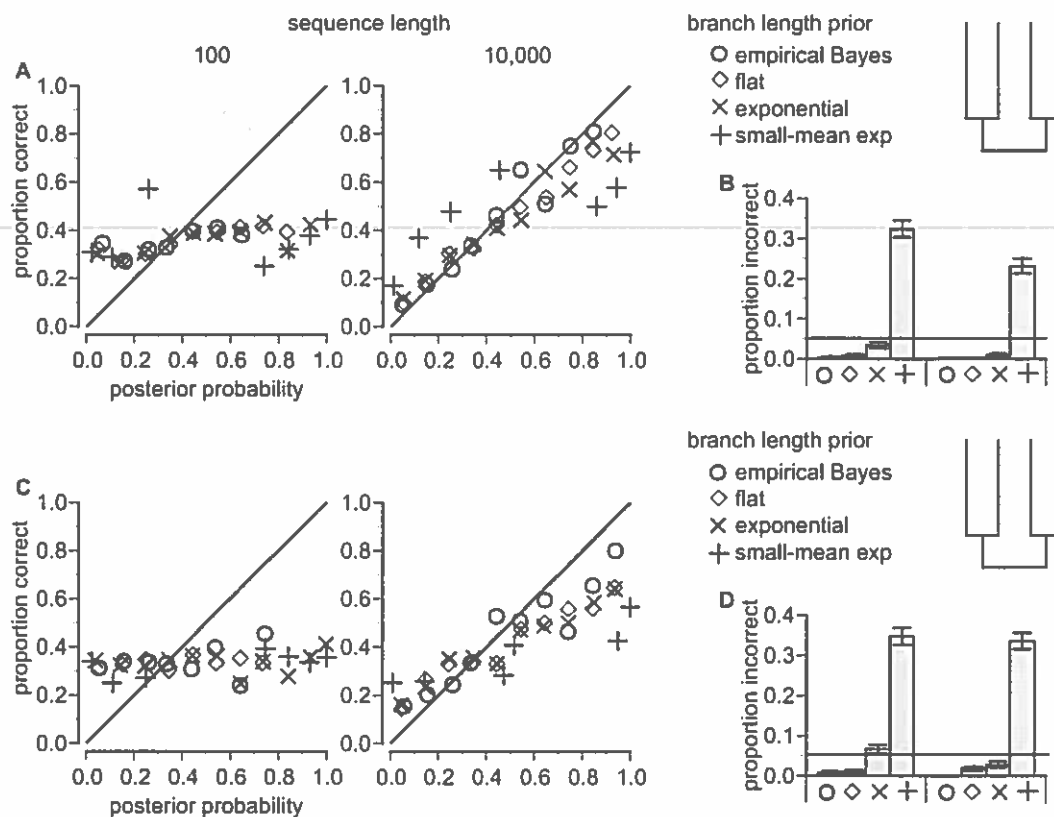
To determine how different branch length patterns affect posterior probabilities when the true branch lengths are not known in advance, we simulated data on four-taxon trees with all possible combinations of long/short terminal branches. We analyzed these data using BMCMC, assuming either a flat branch length prior ( $U(0, 10)$ ), the default prior used in MrBayes v3.1 ( $exp(\mu = 0.1)$ ), or the small-mean exponential prior suggested by Yang and Rannala ( $\mu = 10^{-5}$ ). We additionally calculated posterior probabilities using an empirical Bayes approach that places prior probability 1.0 on the maximum likelihood branch length estimates.

We found that the pattern of branch lengths on the phylogeny used to simulate data can have a strong effect on resulting posterior probabilities. When all terminal branch lengths were short (0.01 substitutions/site), all branch length priors examined produced average posterior probabilities that closely matched the proportion of correct trees (Figure 6.3A), and the rate of false inferences with posterior probability  $\geq 0.95$  was always very low (Figure 6.3B). When only one terminal branch was long (0.75 substitutions/site), all priors except the small-mean exponential produced posterior probabilities that closely matched the proportion of correct trees and low rates of false inferences with strong support (Figure 6.3C-D). In contrast, the small-mean exponential prior produced posterior probabilities greater than the proportion of correct trees when sequences were short (100 nt), resulting in a rate of false inferences significantly greater than 0.05. Longer sequences (10,000 nt) resulted in posterior probabilities close to 1.0 for the correct tree using any branch length prior, and no incorrect trees were resolved with high posterior probability.



**FIGURE 6.3:** Branch length patterns affect posterior probabilities when few terminal branches are long. (A,C) The average posterior probability of a group of inferred trees is plotted against the proportion of correct trees in the group using the binning method; sequences of 1000 and 10,000 nt were examined. (B,D) The proportion of incorrectly-resolved trees with posterior probability  $\geq 0.95$  is shown for various prior distributions; bars indicate standard error, and an error rate of 0.05 is indicated by a horizontal line. Sequence length increases along the horizontal axis, with 1000-nt sequences shown at left and 10,000-nt sequences at right. Panels A,B show results for sequences generated on a tree with a short internal branch (0.01 substitutions/site) and all short terminal branches (0.01), while C,D show results when one of the four terminal branches is long (0.75). Branch length priors examined were an empirical Bayes prior that places prior probability 1.0 on the maximum likelihood branch length estimates (open circles), a flat prior with uniform probability over branch lengths from zero to ten substitutions/site (open diamonds), an exponential prior with mean 0.1 (X's) and a small-mean exponential prior ( $\mu = 10^{-5}$ , crosses).





**FIGURE 6.4:** Branch length patterns affect posterior probabilities when many terminal branches are long. We plotted the average posterior probability of a group of inferred trees against the proportion of correct trees in the group (A,C) and the proportion of trees falsely resolved with posterior probability  $\geq 0.95$  (B,D) when sequences were generated on four-taxon phylogenies with short internal branches (0.01 substitutions/site) and either three (top) or all four (bottom) long terminal branches (0.75). Results for sequences of 1000 nt are shown at the left in each panel, while results for sequences of 10,000 nt are shown at right. Branch length priors examined were the empirical Bayes prior (open circles), a flat prior uniform over (0,10] (open diamonds), an exponential prior with mean 0.1 (X's) and a small-mean exponential prior with mean  $10^{-5}$  (crosses).

Differences between the small-mean exponential and the other branch length prior distributions examined were excentuated when either three or all four terminal branches were long (Figure 6.4). When three terminal branches were long, short sequences produced average posterior probabilities much greater than the proportion of correct trees using any prior distribution; longer sequences lessened

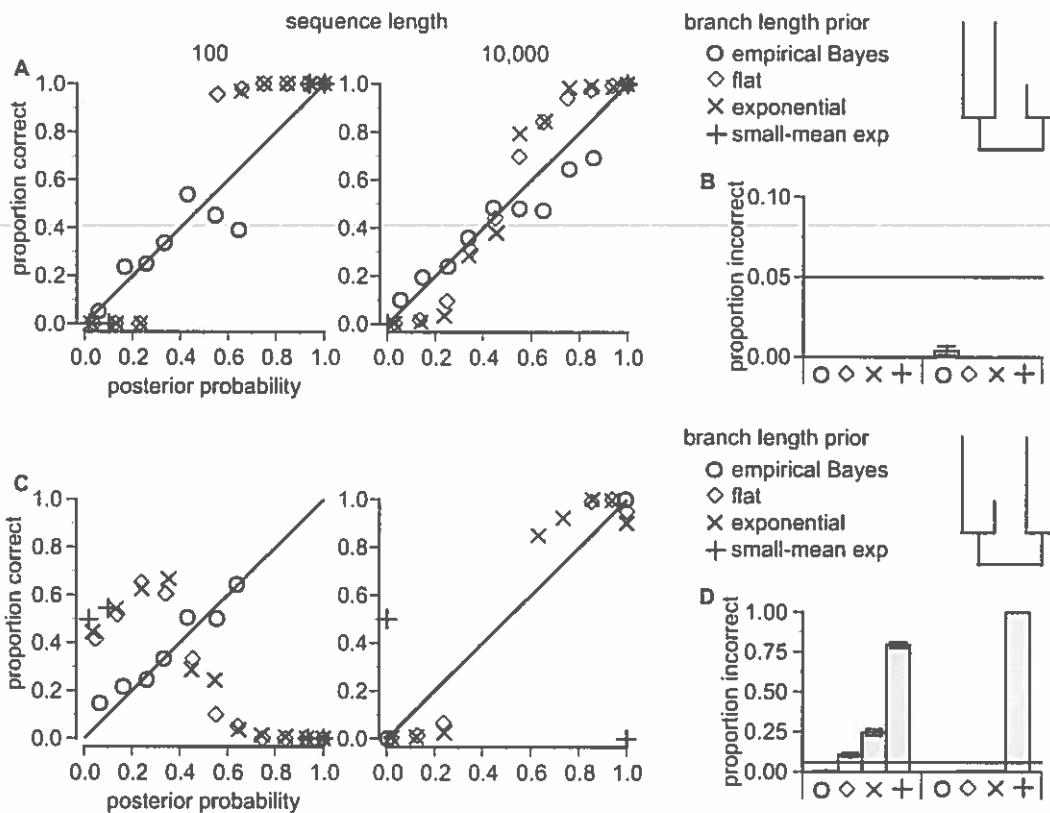
this effect (Figure 6.4A). When sequences were 10,000 nt, posterior probabilities calculated using the empirical Bayes prior were very close to the proportion of correct trees, and those calculated using the flat or exponential prior with moderate mean (0.1) were not much higher. In contrast, average posterior probabilities calculated using the small-mean exponential prior were much higher than the proportion of correct trees. False inference rates were generally low ( $< 0.05$  at posterior probability cutoff 0.95), except the small-mean exponential prior produced excessive rates of false inferences at both sequence lengths examined (Figure 6.4B). Results were similar when all four terminal branches were long (Figure 6.4C-D). Average posterior probabilities were greater than the proportion of correct trees when sequences were short. Longer sequences reduced this effect, with empirical Bayes priors producing posterior probabilities that more closely matched the proportion of correct trees than those produced by other distributions (Figure 6.4C). Posterior probabilities calculated using small-mean exponential priors were substantially greater than those calculated using the other priors, and false inference rates were higher when small-mean exponential priors were used (Figure 6.4D).

When only two terminal branches were long, the relationship between taxa with long branches had a pronounced effect on posterior probabilities (Figure 6.5). When sister taxa had long terminal branches, the empirical Bayes prior produced low posterior probabilities when sequences were short and posterior probabilities slightly higher than the proportion of correct trees when sequences were longer (Figure 6.5A). In contrast, posterior probabilities calculated when branch length uncertainty was integrated over using the other prior distributions were lower than the proportion of correct trees. False inference rates were always low under these conditions (Figure 6.5B). When long terminal branches were not sister to one another, the empirical Bayes prior produced posterior probabilities that closely matched the proportion of correct trees (Figure 6.5C), and false inference rates were low using the empirical Bayes approach (Figure 6.5D). In contrast, integrating over branch length uncertainty resulted in a long-branch attraction artifact with excessive support for an incorrect tree. Short sequences resulted in frequent support for the incorrect tree placing long terminal branches as sister to one another. When

sequences were long, the small-mean exponential prior always inferred the long-branch attraction tree with strong support; the other diffuse priors tended to recover the correct tree, although average posterior probabilities were lower than the proportion of correct inferences (Figure 6.5C). False inference rates were significantly greater than 0.05 (at 0.95 posterior probability cutoff) when sequences were short and branch length uncertainty was integrated over using non-point prior distributions (Figure 6.5D). Strongly supported false trees were absent when sequences were long, except the small-mean exponential prior always recovered the long-branch attraction tree with posterior probability 1.0.

Several general inferences can be drawn from the results of these branch length studies (Figs. 6.2-6.4). First, the small-mean exponential prior produced more extreme posterior probabilities than the other priors examined, resulting in average posterior probabilities with stronger deviations from the proportion of correct trees and a greater frequency of incorrect inferences with high support. Second, longer sequences resulted in average posterior probabilities that more closely matched the proportion of correct trees under all conditions, presumably because of the reduction in branch length uncertainty associated with longer sequences. Longer sequences cause the likelihood function over branch lengths to be more narrowly peaked around the true values, so integrating over that function more closely approximates knowing the true values in advance.

Third, the amount and structure of convergent evolution expected on the true tree appears to determine whether average posterior probabilities will be higher or lower than the proportion of correct trees when branch length uncertainty is integrated over using a non-point prior distribution. Specifically, when one or none of the four terminal branches were long—making convergent state patterns rare—inferred posterior probabilities were close to the proportion of correct inferences (Figure 6.3). When there was ample opportunity for convergent evolution but no expected structure to the convergence—that is, when three or four of the terminal branches were long—average posterior probabilities were higher than the proportion of correct trees (Figure 6.4). In contrast, when convergence was structured to favor a particular tree—as on trees with two long and two short



**FIGURE 6.5:** Branch length patterns affect posterior probabilities when two terminal branches are long (0.75 substitutions/site) and the other two are short (0.01). Sequences of lengths 1000 (left) and 10,000 nt (right) were generated using inverse-Felsenstein zone trees with two long sister branches and two short sister branches (top) and Felsenstein zone trees with non-sister long branches (bottom). We plotted the average posterior probability of a group of trees against the proportion of correct trees in the group using the binning method (A,C) as well as the proportion of trees falsely-resolved with posterior probability  $\geq 0.95$  (B,D) when different branch length priors were assumed. Priors examined were the empirical Bayes prior placing probability 1.0 on the maximum likelihood branch lengths (open circles), a uniform prior over (0,10] (open diamonds), an exponential prior with mean 0.1 (X's), and a small-mean exponential prior with mean  $10^{-5}$  (crosses).

branches—posterior probabilities were lower than the proportion of correct trees (Figure 6.5). When the true tree was in the inverse-Felsenstein zone (two sister long branches), the true tree was always recovered, but posterior probabilities were typically  $< 1.0$  (Figure 6.5A). When the true tree was in the Felsenstein zone (two non-sister long branches), integrating over incorrect branch lengths resulted in a

long-branch attraction bias that reduced support for the correct tree while inflating support for an incorrect tree (Figure 6.5C).

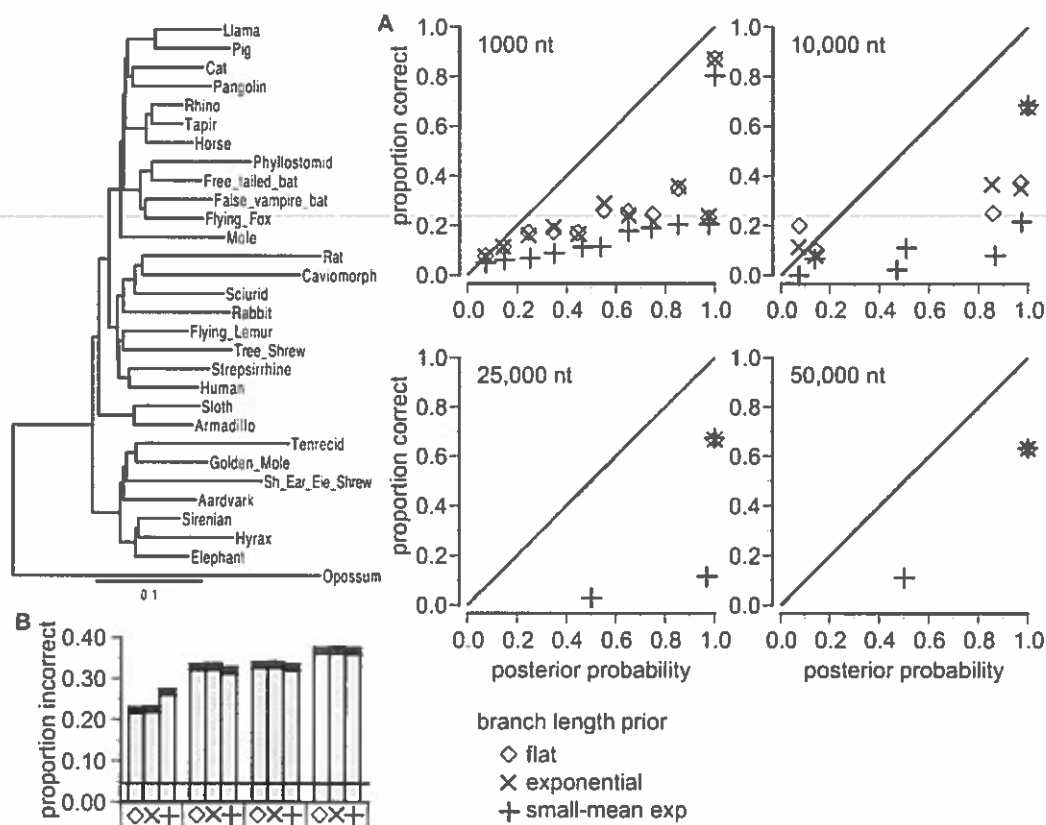
Fourth, we found that either flat or moderate-mean exponential branch length priors consistently yielded posterior probabilities that more closely matched those that would be inferred if the true branch lengths were known in advance than the small-mean exponential prior. Under all conditions that resulted in average posterior probabilities that were higher or lower than the proportion of correct trees, the effect was less severe with the diffuse priors than the small-mean exponential one, and increasing the amount of data reduced the deviation between posterior probability and proportion correct to a greater degree with the flat or moderate-mean exponential priors than the small-mean exponential. For example, when sequences of 10,000 nucleotides were simulated along a tree with four long terminal branches (Figure 6.4A-B), trees with posterior probability  $> 0.94$  using a flat or moderate exponential prior had a 64% chance of being correct, but only 50% of trees with posterior probability  $> 0.94$  using the small-mean exponential prior were correct. Additionally, Felsenstein zone conditions caused long-branch attraction when a small-mean exponential prior was used, resulting in strong support for an incorrect tree even when sequences were very long (Figure 6.5C-D). The flat and moderate exponential priors produced a less severe version of this same effect: short sequences resulted in long-branch attraction; longer sequences allowed inference of the correct tree, although the average posterior probability of the inferred tree was slightly lower than the proportion of trees that were correct.

Finally, the empirical Bayes approach that avoids integration over multiple branch length values by fixing branch lengths at their maximum likelihood estimates produced posterior probabilities more closely matching those that would be inferred if the true branch lengths were known in advance than any of the non-point priors examined. Long branch attraction was completely absent when the empirical Bayes prior was used, and the rate of false inferences was consistently lower than that incurred when branch length uncertainty was incorporated by integration.

### 6.3.4 Even Very Long Sequences Do Not Eliminate the Effects of Branch Length Uncertainty

It is well appreciated that the choice of prior distributions can affect posterior probabilities, and our results above indicate that posterior probabilities for phylogenetic hypotheses may be sensitive to different branch length priors, although they appear to be fairly robust to using either flat or moderate-mean exponential distributions. One potential method for dealing with prior sensitivity is to collect more data. Assuming that the evolutionary model is correct, parameter uncertainty decreases with increasing sequence length: the likelihood function becomes increasingly peaked around the true parameter values as sequences become longer, and the effects of prior assumptions on the posterior distribution will disappear. Although increasing sequence length generally reduced the degree to which average posterior probabilities deviated from the proportion of correct trees in our four-taxon simulations, even relatively long sequences (10,000 nt) were not sufficient to produce posterior probabilities that matched the proportion of correct trees when branch lengths were extreme. This naturally leads to the practical question: how long must sequences be to eliminate the effects of branch length uncertainty on posterior probabilities for real-world problems?

To address this issue, we simulated replicate data sets using parameter values inferred from empirical data, a 30-taxon subset of the placental mammal data of [79]. We estimated the tree, branch lengths, and transition model parameters using the original 16,397-nt alignment; we then simulated sequences of increasing length under these conditions and analyzed them with BMCMC using the correct evolutionary model. We calculated posterior probabilities using the three different branch length priors examined in our four-taxon simulations ( $U(0, 10)$ ,  $exp(0.1)$ , and  $exp(10^{-5})$ ), and average clade probabilities from each prior were compared to the proportion of clades that were correct using the binning method. We also assessed the proportion of sampled clades falsely resolved with posterior probability  $\geq 0.95$  using each branch length prior.



**FIGURE 6.6:** Branch length uncertainty affects posterior probabilities under empirically derived conditions. Sequences were simulated on the tree at left under a complex model with branch lengths and parameters derived from the large mammalian data set of [79] and analyzed by BMCMC using the true evolutionary model with three different branch length priors: 1) a flat prior on branch lengths (open diamonds), 2) an exponential prior with mean 0.1 (X's), and 3) a small-mean exponential prior ( $\mu = 10^{-5}$ , crosses). A) All clades sampled using each prior were binned by their posterior probability, and the fraction of correct clades in each bin was calculated. B) The proportion of sampled clades falsely resolved with posterior probability  $\geq 0.95$  is shown for each branch length prior. Sequence length increases from 1000 to 50,000 nt along the horizontal axis. Bars indicate standard error, and the horizontal line indicates an error rate of 0.05.

As sequence length increased, posterior probabilities for inferred clades converged to 1.0 for all priors examined, although the flat and moderate-mean exponential priors covered faster than small-mean exponential priors in this case (Figure 6.6A). Even though the choice of branch length prior became less important

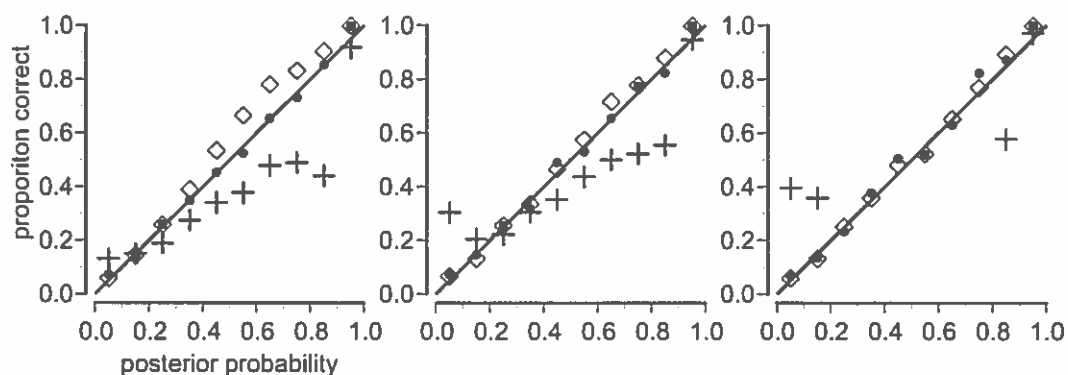
as sequences grew longer, average posterior probabilities were higher than proportion of correct clades for all sequence lengths examined, including 50,000 nucleotides. When sequences were  $\leq 10,000$  nt, the small-mean exponential prior produced higher posterior probabilities than flat or moderate-mean exponential priors. Very long sequences (25,000–50,000 nt) produced posterior probability 1.0 for all inferred clades using the flat and moderate exponential priors, with only 67% of these inferences being correct. The same proportion of inferences with posterior probability 1.0 were correct using the small-mean exponential prior under these conditions, and an additional small number of mostly incorrect clades were more weakly supported. False inference rates were significantly greater than 0.05 at a 0.95 posterior probability cutoff for all branch length priors examined (Figure 6.6B). The underlying mammalian tree for this experiment has very short internal branch lengths and longer terminals, precisely the conditions found above to cause average posterior probabilities to be higher than the proportion of correct inferences in four-taxon simulations. These results suggest that, for difficult real-world problems, extremely long sequences may be required for Bayesian methods to reliably recover the correct tree with high posterior probability, although the effects of different branch length priors on posterior probabilities may disappear at shorter sequence lengths.

### 6.3.5 Diffuse Priors Are Preferable to Small-Mean Exponential Priors

The results described above consistently demonstrate that flat or moderate-mean exponential branch length priors produce average posterior probabilities that more closely match the proportion of correctly-inferred trees than the exponential priors with small mean recommended by Yang and Rannala. Our simulations differ from theirs in two crucial ways. First, we used the standard implementation of a single prior distribution for all branch lengths on the tree, whereas they applied separate priors to terminal and internal branches. Second, we simulated sequences on trees with fixed branch lengths rather than variable lengths drawn from a distribution.



To determine which factor is responsible for the differences in our results, we simulated sequences on ten-taxon phylogenies with branch lengths drawn from an exponential distribution with mean 0.15 substitutions/site and analyzed the data using the true prior ( $exp(0.15)$ ), a flat prior ( $U(0, 10)$ ), or a small-mean exponential prior ( $exp(10^{-5})$ ) applied to all branches of the tree (Figure 6.7). As expected, the true prior produced average posterior probabilities that precisely matched the proportion of correct clades. Posterior probabilities calculated using the flat branch length prior matched the proportion of correct clades at all sequence lengths except 25 nucleotides, at which average posterior probabilities were slightly lower than the proportion of correct clades. In contrast, the small-mean exponential prior produced average clade probabilities higher than the proportion of correct clades at all sequence lengths. These results indicate that a uniform branch length prior produces posterior probabilities that more closely approximate those that would be inferred if the true branch lengths were known in advance than a small-mean exponential prior, whether the true tree has fixed branch lengths or lengths drawn from an unknown distribution.



**FIGURE 6.7:** Small-mean exponential branch length priors produce posterior probabilities that deviate more strongly from those produced using the true prior distribution than uniform priors. Sequences of length 25 (left), 100 (middle), and 1000 nt (right) were simulated on randomly-selected ten-taxon phylogenies with branch lengths drawn from an exponential distribution with  $\mu = 0.15$ . The average posterior probability of a group of clades is plotted against the proportion of correct clades in the group when using the true branch length prior distribution ( $exp(0.15)$ , black dots), a uniform distribution ( $U(0, 10)$ , open diamonds), or an exponential distribution with  $\mu = 10^{-5}$  (crosses).

## 6.4 Discussion

Bayesian phylogenetics is appealing for at least three reasons: 1) BMCMC algorithms allow very large phylogenies to be estimated using complex evolutionary models in a reasonable amount of time; 2) uncertainty in model parameters can be incorporated by integrating over multiple values, and 3) posterior probabilities provide an apparently meaningful measure of statistical confidence in inferred clades. Bayesian techniques have resolved previously intractable problems with strong support [58, 79], but acceptance of these results has been hampered by a growing suspicion that posterior probabilities may regularly be inflated.

Posterior probability is defined by Bayes' Theorem as the probability that a hypothesis is correct given the data, a model of the data-generating process, and prior probability distributions over model parameters. Contrary to early theoretical claims [19], posterior probabilities are not expected to be equivalent to bootstrap proportions, because the two quantities are fundamentally different measures of statistical confidence [3]. Bootstrap proportions attempt to estimate the proportion of replicate data sets that would favor the same tree as the the original data, whereas posterior probabilities measure the support for the favored tree given the data at hand and prior assumptions about the potential values of model parameters. It is also not correct to expect posterior probabilities to always match the proportion of correctly inferred trees, again because posterior probabilities are calculated from the observed data without replication and are conditional on prior assumptions, whereas the proportion of correct inferences necessarily requires replication and is not conditioned on prior assumptions. Nevertheless, our results confirm previous findings that, under computer simulations where replication is possible and when the true values of nuisance parameters are known in advance, the average posterior probability of a group of trees is equivalent to the proporiton of correct trees in the group [53, 132].

In practice, the true values of nuisance parameters such as branch lengths are never known with certainty; our results show that prior uncertainty about the values of branch lengths can affect resulting posterior probabilities. When branch

length values are not known in advance, this uncertainty can either be eliminated by fixing branch lengths at some specified values or incorporated by integrating over multiple lengths using non-point prior distributions. Our results indicate that fixing branch lengths at their maximum likelihood values using an empirical Bayes approach produces posterior probabilities that more closely match those that would be inferred given perfect prior knowledge and result in lower rates of strongly-supported false inferences than integrating over branch lengths using common prior distributions. When branch lengths are not known in advance, integrating over prior uncertainty can dramatically affect posterior probabilities, with both the magnitude and direction of the effect depending on the pattern of branch lengths on the true tree as well as sequence length and the specific prior distributions applied. Posterior probabilities calculated using the empirical Bayes approach are also affected by prior uncertainty, but to a lesser degree. We have shown these results in the case of branch lengths; similar results presumably hold for other evolutionary model parameters as well.

When integrating over a range of plausible values using BMCMC, most of the branch lengths used to calculate the marginal likelihood of a tree will be wrong. The net effect of assuming incorrect branch lengths is to cause convergent state patterns to be misinterpreted as phylogenetic signal or vice versa. With short sequences, the likelihood function is relatively flat over a range of branch lengths, so incorrect lengths contribute substantially to the total likelihood, resulting in strong effects on posterior probabilities. As sequences become longer, the likelihood function becomes more narrowly peaked around the true branch lengths, prior uncertainty decreases, and posterior probabilities become increasingly similar to those that would be calculated given perfect prior knowledge. Under challenging conditions derived from real data, however, we found that posterior probabilities estimated using non-point prior distributions on branch lengths were not the same as those that would be calculated if the true branch lengths were known in advance, even with very long sequences (50,000 nt). The reason for this result is not entirely clear; it could be that the prior uncertainty associated with 50,000-nt sequences is sufficient to affect posterior probabilities when the tree is large and the evolutionary model complex.

An alternative explanation is that because uniform priors over tree topologies do not imply uniform priors over clades, summarizing posterior probabilities on clades over multiple trees could affect posterior probabilities [86]. Our results examining clade probabilities on ten-taxon trees suggest that the summarization process is not a cause of skewed posterior probabilities. Yet another possible explanation is that existing BMCMC algorithms—although sufficient to reliably estimate posterior probabilities for small trees—break down as the inference problem becomes more complicated. Determining the convergence and mixing properties of BMCMC when applied to complex phylogenetic problems would help resolve this issue.

We have shown that—contrary to the recommendation of Yang and Rannala [132]—diffuse branch length priors produce more reliable inferences than exponential priors with very small means across a range of phylogenetic problems. Although Yang and Rannala found that a prior favoring an extremely short internal branch results in low posterior probabilities when the true lengths of terminal branches are known in advance, we have shown that such short branch-length priors, if applied across the entire tree, result in extreme posterior probabilities that deviate from the proportion of correct inferences more severely than those produced using diffuse priors, result in higher rates of false inferences, and are subject to strong long-branch attraction artifacts. These effects are produced because assuming small-mean priors on all branches disfavors substitution in general, resulting in underestimation of convergence when terminal branches are long. In principle, it might be possible to implement a partitioned prior like the one used by Yang and Rannala, where different priors are applied to internal and terminal branches; this approach effectively increases the relative prior probability of convergence, resulting in lower posterior probabilities [132]. Since prior distributions ideally reflect the actual prior beliefs of the investigator, the ability to specify a variety of prior distributions is an important feature to support in Bayesian phylogenetics software.

Uncertainty about nuisance parameters is a critical concern in phylogenetics, because the data themselves are never adequate to precisely specify parameter values with absolute certainty, and different parameter values can produce different results. Maximum likelihood analysis circumvents this issue by fixing nuisance parameters

at their 'best guess' estimates; post-hoc analyses can be used to test the robustness of inferences to parameter uncertainty but require additional resources. In contrast, Bayesian methods formally incorporate parameter uncertainty by integrating over multiple values; the advantage of this approach is that the posterior distribution over parameter values is fully described, but results may be sensitive to prior assumptions that can vary from researcher to researcher. Since prior assumptions affect the posterior distribution, knowing the prior distributions assumed for an analysis—and characterizing the posterior distribution over different prior assumptions—is crucial for interpreting results obtained using Bayesian methods.

There are approaches to characterizing evidentiary support for a clade of interest that are not conditioned on prior knowledge about nuisance parameters (and do not require bootstrapping), such as the likelihood ratio of the best tree with a clade versus the best tree without it [18] and the maximum probability that a clade is false given the data [6]. The empirical Bayes approach examined here is another potentially useful approach to estimating statistical confidence that eliminates the need to specify prior distributions, although the computational demands required to apply the method to large problems may be prohibitively high. Understanding the statistical properties of these and other confidence measures under a variety of conditions warrants further study. Since we feel that no single measure is likely to provide a complete and accurate estimate of statistical confidence under all evolutionary conditions, a careful and critical application of a variety of measures—each evaluated in light of a detailed understanding of its statistical properties—will provide the most robust and thorough assessments of confidence in phylogenetic hypotheses.

## CHAPTER VII

### CONCLUSION

All scientific inference methods make assumptions about the hidden processes from which observed data are generated, and phylogenetic inference is no exception. Current state-of-the-art phylogenetic methods rely on complex molecular evolutionary models describing how sequences change over time. Although existing evolutionary models incorporate many features of molecular evolution, they largely ignore site-specific dynamics in order to maximize the amount of data that can be used to estimate parameters of the model. I have shown that failing to incorporate important site-specific evolutionary dynamics can lead to erroneous inferences. In particular, site-specific changes in evolutionary rates—which have been shown to regularly occur in real molecular sequence data [27, 48, 55, 71, 72, 75, 76, 84, 87, 91]—can confound existing evolutionary models, producing strong support for incorrect phylogenies. I have developed, implemented, and tested a mixed branch length strategy for incorporating heterotachy, showing that it can produce more accurate phylogenetic inferences than existing models under both simulated and real-world conditions. The potential of this model to improve the quality of phylogenetic inferences should be valuable to the biological community, as nearly all biological results are interpretable only in the context of evolutionary history.

The advent of Bayesian phylogenetics is arguably the most important advancement in phylogenetics methodology since the development of model-based methods. The posterior probability of a tree or node—i.e. the probability that the

tree or node is correct given the data [53]—is exactly what we would like a phylogenetic inference method to tell us, and the efficiency of the MCMC algorithm allows posterior probabilities to be calculated on very large phylogenies using complex evolutionary models. Although Bayesian methods have resolved previously intractable problems with strong support, acceptance of these results has been hampered by a number of studies suggesting that posterior probabilities may regularly be too high, resulting in an inflated sense of statistical confidence and a high rate of false inferences. Understanding if, when, and why posterior probabilities are inflated is crucial for interpreting posterior probabilities presented as statistical support for phylogenetic relationships.

I have shown that one of the main proposed causes of inflated posterior probabilities, the star tree paradox, does not actually cause posterior probabilities to be inaccurate. Even when the true tree is unresolved, posterior probabilities calculated using existing algorithms that do not sample the true unresolved tree can provide an accurate measure of statistical confidence. On the other hand, conditioning on incomplete prior knowledge about the values of the evolutionary model's parameters can affect posterior probabilities. When the model's parameter values are not known in advance, Bayesian techniques require prior probability distributions to be placed on all parameters; I have shown that prior uncertainty about branch lengths can cause posterior probabilities to deviate strongly from those that would be inferred given perfect prior knowledge. Different branch length patterns on the true tree can cause posterior probabilities to be skewed either upward or downward when branch length uncertainty is integrated over using diffuse prior distributions. In contrast, an empirical Bayes approach that fixes branch lengths at their maximum likelihood estimates produces posterior probabilities that more closely approximate those that would be inferred if branch lengths were known in advance.

In summary, I have shown that violating key assumptions of phylogenetic inference techniques does make a difference, resulting in the potential for erroneous results and incorrect assessments of statistical confidence in those results. Furthermore, the types of assumption violations I have examined are likely to occur

when real molecular sequence data are analyzed, suggesting that empirical results should be carefully scrutinized to ensure that they are not an artifact of biases induced by assumption violations. In particular, the development of more realistic phylogenetic techniques that relax the simplifying assumptions made by current methods should provide the potential for more accurate results and is an important area for continued research.

These results have important repercussions outside of the biological sciences. Statistical inference techniques are ubiquitous in scientific inquiry, and all statistical inference techniques make assumptions about the unknown processes generating the observed data. In general there is little information available to confirm that these assumptions are correct, and in many cases the assumptions may be wrong. The potential for incorrect inferences due to assumption violations therefore always exists when real data are analyzed to produce inferences. The methodology developed in this dissertation can be used as a general strategy for evaluating the potential accuracy of statistical inference techniques when applied to empirical data. Such an approach requires answering two crucial questions: 1) which assumptions are likely to be violated, and 2) what are the potential effects of such assumption violations? Question 1 must be answered by careful investigations of the system under study. The information gained by these studies can then be used to develop simulation approaches that test the effects of possible assumption violations on existing inference techniques. Understanding how assumption violations effect existing methods can then inform the development of new techniques that better model the system's actual properties and produce more accurate inferences. As scientists, it is crucial that we not only critically evaluate our hypotheses and theories but also the methods we use to evaluate hypotheses and theories.

More generally, whenever automated processes are used to make decisions or infer information from data, assumptions are made about the data that the process observes. Data may not always conform to these assumptions, leading to potentially erroneous results. When designing procedures to automate decisions or control responses to incoming data, it is therefore crucial to investigate the behavior of the decision making procedure when the data violate its assumptions. Robustness to



assumption violations, rather than being a secondary property of automated control systems, is likely to be one of the most important properties for predicting real-world performance.

## BIBLIOGRAPHY

- [1] H. AKAIKE, *A new look at the statistical model identification*, IEEE Transactions on Automatic Control, 19 (1974), pp. 716–723.
- [2] F. AL-AWADHI, M. HURN, AND C. JENNISON, *Improving the acceptance rate of reversible jump mcmc proposals*, Statistics and Probability Letters, 69 (2004), pp. 189–198.
- [3] M. E. ALFARO, S. ZOLLER, AND F. LUTZONI, *Bayes or bootstrap? a simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence*, Molecular Biology and Evolution, 20 (2003), pp. 255–266.
- [4] C. ANDRIEU, N. DE FREITAS, AND A. DOUCET, *Reversible jump MCMC simulated annealing for neural networks*, in Proceedings of the 16th Annual Conference on Uncertainty in Artificial Intelligence, San Francisco, CA, 2000, Morgan Kaufmann, pp. 11–18.
- [5] C. ANE, J. G. BURLEIGH, M. M. MCMAHON, AND M. J. SANDERSON, *Covariation structure in plastid genome evolution: a new statistical test*, Molecular Biology and Evolution, 22 (2005), pp. 914–924.
- [6] M. ANISIMOVA AND O. GASCUEL, *Approximate likelihood ratio test for branches: a fast, accurate and powerful alternative*, Systematic Biology, 55 (2006), pp. 539–552.
- [7] D. BARKER, *LVB: Parsimony and simulated annealing in the search for phylogenetic trees*, Bioinformatics, 20 (2004), pp. 274–275.
- [8] G. BERNARDI, *Isochores and the evolutionary genomics of vertebrates*, Gene, 241 (2000), pp. 3–17.

- [9] H. BRINKMANN, M. V. D. GIEZEN, Y. ZHOU, G. P. DE RAUCORT, AND H. PHILIPPE, *An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics*, *Systematic Biology*, 54 (2005), pp. 743–757.
- [10] T. R. BUCKLEY, *Model misspecification and probabilistic tests of topology: evidence from empirical data sets*, *Systematic Biology*, 51 (2002), pp. 509–523.
- [11] J. T. CHANG, *Inconsistency of evolutionary tree topology reconstruction methods when substitution rates vary across characters*, *Mathematical Biosciences*, 134 (1996), pp. 189–215.
- [12] M. P. CUMMINGS, S. A. HANDLEY, D. S. MYERS, D. L. REED, A. ROKAS, AND K. WINKA, *Comparing bootstrap and posterior probability values in the four-taxon case*, *Systematic Biology*, 52 (2003), pp. 477–487.
- [13] C. DARWIN, *On the Origin of Species*, J Murray, London, 1859.
- [14] C. DELARBRE, C. GALLUT, V. BARRIEL, P. JANVIER, AND G. GACHELIN, *Complete mitochondrial DNA of the hagfish, *Eptatretus burgeri*: the comparative analysis of mitochondrial DNA sequences strongly supports the cyclostome monophyly*, *Molecular Phylogenetics and Evolution*, 22 (2002), pp. 184–192.
- [15] T. DOBZHANSKY, *Nothing in biology makes sense except in the light of evolution*, *American Biology Teacher*, 35 (1973), pp. 25–29.
- [16] T. S. DONALDSON, *Robustness of the  $f$ -test to errors of both kinds and the correlation between the numerator and denominator of the  $f$ -ratio*, *Journal of the American Statistical Association*, 63 (1968), pp. 660–676.
- [17] C. J. DOUADY, F. DELSUC, Y. BOUCHER, W. F. DOOLITTLE, AND E. J. P. DOUZERY, *Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability*, *Molecular Biology and Evolution*, 20 (2003), pp. 248–254.

- [18] A. W. F. EDWARDS, *Likelihood*, Cambridge University Press, Cambridge, Massachusetts, 1972.
- [19] B. EFRON, E. HALLORAN, AND S. HOLMES, *Bootstrap confidence levels for phylogenetic trees*, Proceedings of the National Academy of Sciences USA, 93 (1996), pp. 7085–7090.
- [20] P. ERIXON, B. SVENNBALD, T. BRITTON, AND B. OXELMAN, *Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics*, Systematic Biology, 52 (2003), pp. 665–673.
- [21] J. FELSENSTEIN, *Cases in which parsimony and compatibility methods will be positively misleading*, Systematic Zoology, 27 (1978), pp. 401–410.
- [22] J. FELSENSTEIN, *Evolutionary trees from DNA sequences: A maximum likelihood approach*, Journal of Molecular Evolution, 17 (1981), pp. 368–376.
- [23] J. FELSENSTEIN, *Confidence limits on phylogenies: an approach using the bootstrap*, Evolution, 39 (1985), pp. 783–791.
- [24] J. FELSENSTEIN, *Inferring Phylogenies*, Sinauer Associates Sunderland Massachusetts, 2004.
- [25] N. M. FERGUSON, A. P. GALVANI, AND R. M. BUSH, *Ecological and immunological determinants of influenza evolution*, Nature, 422 (2003), pp. 428–433.
- [26] W. M. FITCH, *The nonidentity of invariable positions in the cytochromes c of different species*, Biochemical Genetics, 5 (1971), pp. 231–241.
- [27] W. M. FITCH, *The molecular evolution of cytochrome c in eukaryotes*, Journal of Molecular Evolution, 8 (1976), pp. 13–40.
- [28] S. R. GADAGKAR AND S. KUMAR, *Maximum likelihood outperforms maximum parsimony even when evolutionary rates are heterotachous*, Molecular Biology and Evolution, 22 (2005), pp. 2139–2141.

- [29] N. GALTIER, *Maximum-likelihood phylogenetic analysis under a covarion-like model*, *Molecular Biology and Evolution*, 18 (2001), pp. 866–873.
- [30] E. A. GAUCHER AND M. M. MIYAMOTO, *A call for likelihood phylogenetics even when the process of sequence evolution is heterogeneous*, *Molecular Phylogenetics and Evolution*, 37 (2005), pp. 928–931.
- [31] E. A. GAUCHER, M. M. MIYAMOTO, AND S. A. BENNER, *Function-structure analysis of proteins using covarion-based evolutionary approaches: elongation factors*, *Proceedings of the National Academy of Sciences USA*, 98 (2001), pp. 548–552.
- [32] B. S. GAUT AND P. O. LEWIS, *Success of maximum likelihood phylogeny inference in the four-taxon case*, *Molecular Biology and Evolution*, 12 (1995), pp. 152–162.
- [33] S. GEMAN AND D. GEMAN, *Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images*, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6 (1984), pp. 721–741.
- [34] B. T. GRENFELL, O. G. PYBUS, J. R. GOG, J. L. N. WOOD, J. M. DALY, J. A. MUMFORD, AND E. C. HOLMES, *Unifying the epidemiological and evolutionary dynamics of pathogens*, *Science*, 303 (2004), pp. 327–332.
- [35] S. GRIBALDO, D. CASANE, P. LOPEZ, AND H. PHILIPPE, *Functional divergence prediction from evolutionary analysis: A case study of vertebrate hemoglobin*, *Molecular Biology and Evolution*, 20 (2003), pp. 1754–1759.
- [36] X. GU, *Functional divergence in protein (family) sequence evolution*, *Genetica*, 118 (2003), pp. 133–141.
- [37] S. GUINDON AND O. GASCUEL, *A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood*, *Systematic Biology*, 52 (2003), pp. 696–704.

- [38] B. HAJEK, *Cooling schedules for optimal annealing*, Mathematics of Operations Research, 13 (1988), pp. 311–329.
- [39] U. H. E. HANSMANN, *Parallel tempering algorithm for conformational studies of biological molecules*, Chemical Physics Letters, 281 (1997), pp. 140–150.
- [40] P. H. HARVEY AND M. D. PAGEL, *The comparative method in evolutionary biology*, Oxford series in ecology and evolution, Oxford University Press, 1991.
- [41] M. HASEGAWA AND M. FUJIWARA, *Relative efficiencies of the maximum likelihood, maximum parsimony, and neighbor-joining methods for estimating protein phylogeny*, Molecular Phylogenetics and Evolution, 2 (1993), pp. 1–5.
- [42] M. HASEGAWA, H. KISHINO, AND T. YANO, *Dating the human-ape splitting by a molecular clock of mitochondrial DNA*, Journal of Molecular Evolution, 22 (1985), pp. 160–174.
- [43] D. M. HILLIS, *Taxonomic sampling, phylogenetic accuracy, and investigator bias*, Systematic Biology, 47 (1998), pp. 3–8.
- [44] D. M. HILLIS AND J. J. BULL, *An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis*, Systematic Biology, 42 (1993), pp. 182–192.
- [45] D. M. HILLIS, J. P. HUELSENBECK, AND C. W. CUNNINGHAM, *Application and accuracy of molecular phylogenies*, Science, 264 (1994), pp. 671–677.
- [46] J. P. HUELSENBECK, *Performance of phylogenetic methods in simulation*, Systematic Biology, 44 (1995), p. 32.
- [47] J. P. HUELSENBECK, *Systematic bias in phylogenetic analysis: Is the strepsiptera problem solved?*, Systematic Biology, 47 (1998), pp. 519–537.
- [48] J. P. HUELSENBECK, *Testing a covarion model of DNA substitution*, Molecular Biology and Evolution, 19 (2002), pp. 689–707.

- [49] J. P. HUELSENBECK AND D. M. HILLIS, *Success of phylogenetic methods in the four-taxon case*, *Systematic Biology*, 42 (1993), pp. 247–264.
- [50] J. P. HUELSENBECK, B. LARGET, AND M. E. ALFARO, *Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo*, *Molecular Biology and Evolution*, 21 (2004), pp. 1123–1133.
- [51] J. P. HUELSENBECK, B. LARGET, R. E. MILLER, AND F. RONQUIST, *Potential applications and pitfalls of Bayesian inference of phylogeny*, *Systematic Biology*, 51 (2002), pp. 673–688.
- [52] J. P. HUELSENBECK AND B. RANNALA, *Phylogenetic methods come of age: Testing hypotheses in an evolutionary context*, *Science*, 276 (1997), pp. 227–232.
- [53] J. P. HUELSENBECK AND B. RANNALA, *Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models*, *Systematic Biology*, 53 (2004), pp. 904–913.
- [54] J. P. HUELSENBECK, F. RONQUIST, R. NIELSEN, AND J. P. BOLLBACK, *Bayesian inference of phylogeny and its impact on evolutionary biology*, *Science*, 294 (2001), pp. 2310–2314.
- [55] Y. INAGAKI, E. SUSKO, N. M. FAST, AND A. J. ROGER, *Covarion shifts cause a long-branch attraction artifact that unites microsporidia and archaeobacteria in EF-1 $\alpha$  phylogenies*, *Molecular Biology and Evolution*, 21 (2004), pp. 1340–1349.
- [56] L. INGBER, *Very fast simulated re-annealing*, *Mathematical and Computer Modelling*, 12 (1989), pp. 967–973.
- [57] D. T. JONES, W. R. TAYLOR, AND J. M. THORNTON, *The rapid generation of mutation data matrices from protein sequences*, *CABIOS*, 8 (1992), pp. 275–282.

- [58] K. G. KAROL, R. M. MCCOURT, M. T. CIMINO, AND C. F. DELWICHE, *The closest living relatives of land plants*, *Science*, 294 (2001), pp. 2351–2353.
- [59] M. KIMURA, *A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences*, *Journal of Molecular Evolution*, 16 (1980), pp. 111–120.
- [60] S. KIRKPATRICK, J. C. D. GELATT, AND M. P. VECCHI, *Optimization by simulated annealing*, *Science*, 220 (1983), pp. 671–680.
- [61] B. KIRKUP AND J. KIM, *From rolling hills to jagged mountains: Scaling of heuristic searches for phylogenetic estimation*, *Molecular Biology and Evolution*, In Review (2006).
- [62] B. KOLACZKOWSKI AND J. W. THORNTON, *Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous*, *Nature*, 431 (2004), pp. 980–984.
- [63] M. K. KUHNER AND J. FELSENSTEIN, *A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates*, *Molecular Biology and Evolution*, 11 (1994), pp. 459–468.
- [64] J. LAM AND J. M. DELOSME, *Performance of a new annealing schedule*, *Proceedings of the 25th ACM/IEEE Design Automation Conference*, (1988), pp. 306–311.
- [65] A. R. LEMMON AND E. C. MORIARTY, *The importance of proper model assumption in Bayesian phylogenetics*, *Systematic Biology*, 53 (2004), pp. 265–277.
- [66] P. O. LEWIS, *NCL: a C++ class library for interpreting data files in NEXUS format*, *Bioinformatics*, 19 (2003), pp. 2330–2331.
- [67] P. O. LEWIS, M. T. HOLDER, AND K. E. HOLSINGER, *Polytomies and Bayesian phylogenetic inference*, *Systematic Biology*, 54 (2005), pp. 241–253.



- [68] P. LOCKHART, A. W. D. LARKUM, M. STEEL, P. J. WADDELL, AND D. PENNY, *Evolution of chlorophyll and bacteriochlorophyll: The problem of invariant sites in sequence analysis*, Proceedings of the National Academy of Sciences USA, 93 (1996), pp. 1930–1934.
- [69] P. LOCKHART, P. NOVIS, B. G. MILLIGAN, J. RIDEN, A. RAMBAUT, AND T. LARKUM, *Heterotachy and tree building: a case study with plastids and eubacteria*, Molecular Biology and Evolution, 23 (2006), pp. 40–45.
- [70] P. J. LOCKHART AND M. STEEL, *A tale of two processes*, Systematic Biology, 54 (2005), pp. 948–951.
- [71] P. J. LOCKHART, M. A. STEEL, A. C. BARBROOK, D. H. HUSON, M. A. CHARLESTON, AND C. J. HOWE, *A covariotide model explains apparent phylogenetic structure of oxygenic photosynthetic lineages*, Molecular Biology and Evolution, 15 (1998), pp. 1183–1188.
- [72] P. LOPEZ, D. CASANE, AND H. PHILIPPE, *Heterotachy, an important process of protein evolution*, Molecular Biology and Evolution, 19 (2002), pp. 1–7.
- [73] G. MCLACHLAN AND D. PEEL, *Finite Mixture Models*, Wiley Interscience, New York, 2000.
- [74] K. MISAWA AND M. NEI, *Reanalysis of Murphy et al.'s data gives various mammalian phylogenies and suggests overcredibility of Bayesian trees*, Journal of Molecular Evolution, 57 (2003), pp. S290–S296.
- [75] B. MISOF, C. L. ANDERSON, T. R. BUCKLEY, D. ERPENBECK, A. RICKERT, AND K. MISOF, *An empirical analysis of MT 16S rRNA covarion-like evolution in insects: site-specific rate variation is clustered and frequently detected*, Journal of Molecular Evolution, 55 (2002), pp. 460–469.
- [76] M. M. MIYAMOTO AND W. M. FITCH, *Testing the covarion hypothesis of molecular evolution*, Molecular Biology and Evolution, 12 (1995), pp. 503–513.

- [77] D. MOREIRA, S. KERVESTIN, O. JEAN-JEAN, AND H. PHILIPPE, *Evolution of eukaryotic translation elongation and termination factors: variations of evolutionary rate and genetic code deviations*, *Molecular Biology and Evolution*, 19 (2002), pp. 189–200.
- [78] E. MOSSEL AND E. VIGODA, *Phylogenetic MCMC algorithms are misleading on mixtures of trees*, *Science*, 309 (2005), pp. 2207–2209.
- [79] W. J. MURPHY, E. EIZIRIK, S. J. O'BRIEN, O. MADSEN, M. SCALLY, C. J. DOUADY, E. TEELING, O. A. RYDER, M. J. STANHOPE, W. W. DE JONG, AND M. S. SPRINGER, *Resolution of the early placental mammal radiation using Bayesian phylogenetics*, *Science*, 294 (2001), pp. 2348–2351.
- [80] G. J. NAYLOR AND W. M. BROWN, *Amphioxus mitochondrial DNA, chordate phylogeny, and the limits of inference based on comparisons of sequences*, *Systematic Biology*, 47 (1998), pp. 61–76.
- [81] H. PHILIPPE, D. CASANE, S. GRIBALDO, P. LOPEZ, AND J. MEUNIER, *Heterotachy and functional shift in protein evolution*, *International Union of Biochemistry and Molecular Biology: Life*, 55 (2003), pp. 257–265.
- [82] H. PHILIPPE AND A. GERMOT, *Phylogeny of eukaryotes based on ribosomal RNA: Long-branch attraction and models of sequence evolution*, *Molecular Biology and Evolution*, 17 (2000), pp. 830–834.
- [83] H. PHILIPPE, N. LARTILLOT, AND H. BRINKMANN, *Multigene analyses of bilaterian animals corroborate the monophyly of ecdysozoa, lophotrochozoa, and protostomia*, *Molecular Biology and Evolution*, 22 (2005), pp. 1246–1253.
- [84] H. PHILIPPE AND P. LOPEZ, *On the conservation of protein sequences in evolution*, *Trends in Biochemical Sciences*, 26 (2001), pp. 4–4–416.
- [85] H. PHILIPPE, Y. ZHOU, H. BRINKMANN, N. RODRIGUE, AND F. DELSUC, *Heterotachy and long-branch attraction in phylogenetics*, *BMC Evolutionary Biology*, 5 (2005).

- [86] K. M. PICKETT AND C. P. RANDLE, *Strange Bayes indeed: uniform topological priors imply non-uniform clade priors*, *Molecular Phylogenetics and Evolution*, 34 (2005), pp. 203–211.
- [87] D. D. POLLOCK, W. R. TAYLOR, AND N. GOLDMAN, *Coevolving protein residues: maximum likelihood identification and relationship to structure*, *Journal of Molecular Biology*, 287 (1999), pp. 187–198.
- [88] D. POSADA AND T. R. BUCKLEY, *Model selection and model averaging in phylogenetics: Advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests*, *Systematic Biology*, 53 (2004), pp. 793–808.
- [89] D. POSADA AND K. A. CRANDALL, *Modeltest: testing the model of DNA substitution*, *Bioinformatics*, 14 (1998), pp. 817–818.
- [90] D. POSADA AND K. A. CRANDALL, *Selecting the best-fit model of nucleotide substitution*, *Systematic Biology*, 50 (2001), pp. 580–601.
- [91] T. PUPKO AND N. GALTIER, *A covarion-based method for detecting molecular adaptation: application to the evolution of primate mitochondrial genomes*, *Proceedings of the Royal Society B: Biological Sciences*, 269 (2002), pp. 1313–1316.
- [92] B. RANNALA AND Z. YANG, *Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference*, *Journal of Molecular Evolution*, 43 (1996), pp. 304–311.
- [93] G. O. ROBERTS AND J. S. ROSENTHAL, *Markov chain Monte Carlo: some practical implications of theoretical results*, *Canadian Journal of Statistics*, 26 (1998), pp. 5–31.
- [94] A. ROKAS, B. L. WILLIAM, N. KING, AND S. B. CARROLL, *Genome-scale approaches to resolving incongruence in molecular phylogenies*, *Nature*, 425 (2003), pp. 798–804.

- [95] F. RONQUIST AND J. P. HUELSENBECK, *MrBayes 3: Bayesian phylogenetic inference under mixed models*, *Bioinformatics*, 19 (2003), pp. 1572–1574.
- [96] C. A. RUSSO, N. TAKEZAKI, AND M. NEI, *Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny*, *Molecular Biology and Evolution*, 13 (1996), pp. 525–536.
- [97] L. A. SALTER AND D. K. PEARL, *Stochastic search strategy or estimation of maximum likelihood phylogenetic trees*, *Systematic Biology*, 50 (2001), pp. 7–17.
- [98] M. J. SANDERSON AND J. KIM, *Parametric phylogenetics?*, *Systematic Biology*, 49 (2000), pp. 817–829.
- [99] H. SHIMODAIRA, *An approximately unbiased test of phylogenetic tree selection*, *Systematic Biology*, 51 (2002), pp. 492–508.
- [100] H. SHIMODAIRA AND M. HASEGAWA, *CONSEL: for assessing the confidence of phylogenetic tree selection*, *Bioinformatics*, 17 (2001), pp. 1246–1247.
- [101] M. SIDDALL AND A. KLUDGE, *Letter to the editor*, *Cladistics*, 15 (1999), p. 2.
- [102] M. P. SIMMONS, K. M. PICKETT, AND M. MIYA, *How meaningful are Bayesian support values?*, *Molecular Biology and Evolution*, 21 (2004), pp. 188–199.
- [103] T. SITNIKOVA, A. RZHETSKY, AND M. NEI, *Interior-branch and bootstrap tests of phylogenetic trees*, *Molecular Biology and Evolution*, 12 (1995), pp. 319–333.
- [104] M. SPENCER, E. SUSKO, AND A. J. ROGER, *Likelihood, parsimony, and heterogeneous evolution*, *Molecular Biology and Evolution*, 22 (2005), pp. 1161–1164.

- [105] A. STAMATAKIS, *An efficient program for phylogenetic inference using simulated annealing*, Proceedings of the 19th IEEE International Parallel and Distributed Processing Symposium, (2005), pp. 198–205.
- [106] A. STAMATAKIS, T. LUDWIG, AND HARALD MEIER, *RAxML-II: A program for sequential, parallel and distributed inference of large phylogenetic trees*, Concurrency and Computation: Practice and Experience, 0 (2003), pp. 1–7.
- [107] M. STEEL, *Should phylogenetic models be trying to 'fit an elephant'?*, Trends in Genetics, 21 (2005), pp. 307–309.
- [108] M. STEEL, D. HUSON, AND P. J. LOCKHART, *Invariable sites models and their use in phylogeny reconstruction*, Systematic Biology, 49 (2000), pp. 225–232.
- [109] J. SULLIVAN AND D. L. SWOFFORD, *Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution pattern are violated?*, Systematic Biology, 50 (2001), pp. 723–729.
- [110] J. SULLIVAN, D. L. SWOFFORD, AND G. J. P. NAYLOR, *The effects of taxon sampling on estimating rate heterogeneity parameters of maximum-likelihood models*, Molecular Biology and Evolution, 6 (1999), p. 10.
- [111] E. SUSKO, M. SPENCER, AND A. J. ROGER, *Biases in phylogenetic estimation can be caused by random sequence segments*, Journal of Molecular Evolution, 61 (2005), pp. 351–359.
- [112] Y. SUZUKI, G. V. GLAZKO, AND M. NEI, *Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics*, Proceedings of the National Academy of Sciences USA, 99 (2002), pp. 16138–16143.
- [113] R. H. SWENDSEN AND J.-S. WANG, *Replica monte carlo simulation of spin-glasses*, Physical Review Letters, 57 (1986), pp. 2607–2609.

- [114] D. L. SWOFFORD, *PAUP\*: Phylogenetic Analysis Using Parsimony and Other Methods, v.4.0b10*, Sinauer Associates, Sunderland, Massachusetts, 1998.
- [115] D. L. SWOFFORD, P. J. WADDELL, J. P. HUELSENBECK, P. G. FOSTER, P. O. LEWIS, AND J. S. ROGERS, *Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods*, *Systematic Biology*, 50 (2001), pp. 525–539.
- [116] Y. TATENO, N. TAKEZAKI, AND M. NEI, *Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site*, *Molecular Biology and Evolution*, 11 (1994), pp. 261–277.
- [117] D. J. TAYLOR AND W. H. PIEL, *An assessment of accuracy, error, and conflict with support values from genome-scale phylogenetic data*, *Molecular Biology and Evolution*, 21 (2004), pp. 1534–1537.
- [118] J. W. THORNTON AND B. KOLACZKOWSKI, *No magic pill for phylogenetic error*, *Trends in Genetics*, 21 (2005), pp. 310–311.
- [119] C. TUFFLEY AND M. STEEL, *Links between maximum likelihood and maximum parsimony under a simple model of site substitution*, *Bulletin of Mathematical Biology*, 59 (1997), pp. 581–607.
- [120] C. TUFFLEY AND M. STEEL, *Modeling the covarion hypothesis of nucleotide substitution*, *Mathematical Biosciences*, 147 (1998), pp. 63–91.
- [121] L. S. VINH AND A. VON HAESLER, *IQPNNI: Moving fast through tree space and stopping in time*, *Molecular Biology and Evolution*, 21 (2004), pp. 1565–1571.
- [122] C. O. WEBB, D. D. ACKERLY, M. A. MCPEEK, AND M. J. DONOGHUE, *Phylogenies and community ecology*, *Annual Review of Ecology and Systematics*, 33 (2002), pp. 475–505.

- [123] S. WHELAN AND N. GOLDMAN, *A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach*, *Molecular Biology and Evolution*, 18 (2001), pp. 691–699.
- [124] T. P. WILCOX, D. J. ZWICKL, T. A. HEATH, AND D. M. HILLIS, *Phylogenetic relationships of the dwarf boas and a comparison of Bayesian and bootstrap measures of phylogenetic support*, *Molecular Phylogenetics and Evolution*, 25 (2002), pp. 361–371.
- [125] E. O. WILSON AND B. HOLLOBLER, *The rise of the ants: A phylogenetic and ecological explanation*, *Proceedings of the National Academy of Sciences USA*, 102 (2005), p. 4.
- [126] Z. YANG, *Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods*, *Journal of Molecular Evolution*, 39 (1994), pp. 306–314.
- [127] Z. YANG, *Evaluation of several methods for estimating phylogenetic trees when substitution rates differ over nucleotide sites*, *Journal of Molecular Evolution*, 39 (1995), p. 9.
- [128] Z. YANG, *Among-site rate variation and its impact on phylogenetic analyses*, *Trends in Ecology and Evolution*, 11 (1996), p. 367.
- [129] Z. YANG, *Maximum-likelihood models for combined analyses of multiple sequence data*, *Journal of Molecular Evolution*, 42 (1996), pp. 587–596.
- [130] Z. YANG, *PAML: a program package for phylogenetic analysis by maximum likelihood*, *Computer Applications in the Biosciences*, 13 (1997), pp. 555–556.
- [131] Z. YANG, N. GOLDMAN, AND A. FRIDAY, *Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation*, *Molecular Biology and Evolution*, 11 (1994), pp. 316–324.
- [132] Z. YANG AND B. RANNALA, *Branch-length prior influences Bayesian posterior probability of phylogeny*, *Systematic Biology*, 54 (2005), pp. 455–470.

- [133] D. J. ZWICKL AND M. T. HOLDER, *Model parameterization, prior distributions, and the general time reversible model in Bayesian phylogenetics*, *Systematic Biology*, 53 (2004), pp. 877–888.



